Evaluación psicológica: Historia, principios y aplicaciones



EL LIBRO MUERE CUANDO LO FOTOCOPIA

AMIGO LECTOR:

La obra que usted tiene en sus manos posee un gran valor.
En ella, su autor ha vertido conocimientos, experiencia y mucho trabajo. El editor ha procurado una presentación digna de su contenido y está poniendo todo su empeño y recursos para que sea ampliamente difundida, a través de su red de comercialización.

Al fotocopiar este libro, el autor y el editor dejan de percibir lo que corresponde a la inversión que ha realizado y se desalienta la creación de nuevas obras. Rechace cualquier ejemplar "pirata" o fotocopia ilegal de este libro, pues de lo contrario estará contribuyendo al lucro de quienes se aprovechan ilegítimamente del esfuerzo del autor y del editor.

La reproducción no autorizada de obras protegidas por el derecho de autor no sólo es un delito, sino que atenta contra la creatividad y la difusión de la cultura.

Para mayor información comuníquese con nosotros:

Manual Moderno®

Editorial El Manual Moderno, S. A. de C. V.
Av. Sonora 206, Col Hipódromo, 06100
México, D.F.

Editorial El Manual Moderno (Colombia), Ltda
Carrera 12-A No. 79-03/05
Bogotá, D.C.

CeMPro
Centro Mexicano de Protección y Fomento a los Derechos de Autor
Sociedad de Gestión Colectiva

Primera Edición En Español
Traducida De La
Tercera Edición En Inglés

Evaluación Psicológica: Historia, principios y aplicaciones

Robert J. Gregory
Wheaton College, Wheaton, Illinois

Traducción según la 3a. edición en inglés por:
Psic. Gloria Padilla Sierra
Facultad de Psicología
UNAM

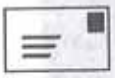
Editor responsable:
Lic. Verónica Ontiveros Ruiz
Editorial El Manual Moderno

Manual Moderno®

Editorial El Manual Moderno, S.A. de C.V.
Av. Sonora 206 Col. Hipódromo, C.P. 06100 México, D.F.

Editorial El Manual Moderno, (Colombia), Ltda
Carrera 12-A No. 79-03/05 Bogotá, DC

Nos interesa su opinión, comuníquese con nosotros:



Editorial El Manual Moderno, S.A. de C.V.,
Av. Sonora núm. 206,
Col. Hipódromo,
Deleg. Cuauhtémoc,
06100 México, D.F.



(52-55)52-65-11-62



(52-55)52-65-11-00



info@manualmoderno.com

Para mayor información sobre:
* Catálogo de producto
* Novedades
* Pruebas psicológicas en línea y más
www.manualmoderno.com

Título original de la obra:
Psychological Testing
History, Principles, and Applications
Copyright © 2000, 1996, 1992
ISBN 0-205-30479-6
Allyn & Bacon, Inc.
A Pearson Education Company
160 Gould Street
Needham Heights, MA 02494

Internet: www.abacon.com

Evaluación psicológica:
historia, principios y aplicaciones
D.R. © 2001 por Editorial El Manual Moderno, S.A. de C.V.
ISBN 968-426-888-2

Miembro de la Cámara Nacional
de la Industria Editorial Mexicana, Reg. núm. 39

Todos los derechos reservados. Ninguna parte de esta publicación puede ser reproducida, almacenada en sistema alguno de tarjetas perforadas o transmitida por otro medio —electrónico, mecánico, fotocopiador, registrador, etcétera— sin permiso previo por escrito de la Editorial.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission in writing from the Publisher.

***Manual Moderno*®**
es marca registrada de
Editorial El Manual Moderno, S.A. de C.V.

Gregory, Robert J., 1943-
Evaluación psicológica : historia, principios y aplicaciones / Robert J. Gregory ; tr. Gloria Padilla Sierra. -- México : Editorial El Manual Moderno, 2001.
xvi, 815 p. : il. ; 24 cm.
Traducción de: Psychological testing : history, principles and applications -- 3rd ed.
Incluye índice
Bibliografía: p. 737-790
ISBN 968-426-888-2
1. Pruebas psicológicas. 2. Pruebas psicológicas - Historia. I. Padilla Sierra, Gloria, tr. II. t.

150.287 GRE.e. 2001 Biblioteca Nacional de México

Reconocimientos

Desearía expresar mi gratitud a diversas personas por ayudar a que la tercera edición en inglés (1a. en español) se volviera una realidad. Sean Wakely y Sue Hutchinson proporcionaron el aliento y consejo en las diversas fases de la revisión. Deborah Brown guió la obra a través de la fase de producción y ayudó a determinar el diseño del libro. Docenas de psicólogos y educadores permitieron reproducir cuadros, figuras e ilustraciones de sus investigaciones y trabajos académicos. Más que reunir estos nombres en un oscuro apéndice que pocos lectores verían, he citado a quienes contribuyeron con el libro dentro de este contexto; además, estas personas ayudaron con las ediciones anteriores, y su orientación ha proseguido en la versión actual:

George M. Alliger, *University of Albany.*
Linda J. Allred, *East Caroline University.*

Kay Bathurst, *California State University,* Fullerton.

Fred Brown, *Iowa State University.*

Michael L. Chase, *Quincy University.*

Milton J. Dehn, *University of Wisconsin-*La Crosse.

Timothy S. Hartshorne, *Central Michigan University.*

Herbert W. Helm, Jr., *Andrews University.*

Ted Jaeger, *Westminster College.*

Richard Kimball, *Worcester State College.*

Haig J. Kojian.

Phyllis M. Ladrigan, *Nazareth College.*

Terry G. Newell, *California State University,* Fresno.

Walter L. Porter, *Harding University.*

Linda Krug Porzelius, *SUNY,* Brockport.

Robert W. Read, *Northeastern University.*

Robert A. Reeves, *Augusta College.*

James R. Sorensen, *Northeastern University.*

Billy Van Jones, *Abilene Christian University.*

También se le debe un agradecimiento a los muchos editores que autorizaron la reproducción de materiales. A los administradores y colegas en *Wheaton College* (Illinois) que proporcionaron recursos excelentes y una atmósfera de apoyo. Tres asistentes de posgrado merecen un agradecimiento especial por encontrar rápidamente las referencias pertinentes: Timothy Chaddock, John Laskowski y Kevin Novotny.

Por último, quiero dar un agradecimiento especial a Mary, Sara y Anne, quienes siguen apoyando mi inquietud por la redacción de textos. Al menos por algunos años prometo no mencionar "el libro" cuando mis seres queridos me pregunten cómo van las cosas.

Prefacio

Las pruebas psicológicas comenzaron como una insignificante empresa en los laboratorios académicos de los psicólogos europeos del siglo XIX. A partir de este nacimiento poco propicio, su práctica proliferó a lo largo de todo el mundo industrializado a un paso cada vez más acelerado. Como descubrirá el lector dentro de las páginas de este libro, las pruebas psicológicas tienen en la actualidad un impacto en casi todos los ámbitos de la vida moderna, desde la educación, pasando por la vocación, hasta la regularización.

PROPÓSITO DEL LIBRO

La tercera edición en inglés (1a. en español) de este libro se basa en las mismas suposiciones que las versiones anteriores. Su ambicioso propósito consiste en proporcionar al lector el conocimiento acerca de las características, objetivos y efectos de amplio rango de la trascendental empresa de las pruebas psicológicas. En la búsqueda de esta meta se han incorporado ciertas tradiciones muy conocidas, pero tam-

bién se ha procedido hacia algunas direcciones nuevas. Por ejemplo, el libro trata el tema común de las normas, estandarización, confiabilidad, validez y construcción de pruebas; también se analiza un compendio de pruebas y medidas en áreas tradicionales como las pruebas intelectuales, de rendimiento, industriales-organizacionales, vocacionales y de personalidad.

CARACTERÍSTICAS ESPECIALES

Además de los temas tradicionales que se listan antes, se enfatizan ciertas cuestiones, temas y conceptos que se consideran esenciales para una comprensión profunda de las pruebas psicológicas; por ejemplo, el libro inicia con un capítulo acerca de la historia de éstas, y subraya la perspectiva de que la historia de las pruebas psicológicas es de importancia para las prácticas actuales. Es típico que los estudiantes de psicología rehuyan las cuestiones históricas debido a que, a menudo, se les presenta de manera aburrida, desprovistas de relevancia para la actualidad. Sin embargo, se espera que el lector escéptico aborde este capítulo con una mente abierta —pues se ha hecho un gran esfuerzo en hacerlo interesante y relevante.

Las pruebas psicológicas representan un contrato entre dos personas. Una de ellas —el examinador— generalmente ocupa una posición de poder por encima de la otra —la persona evaluada. Por tal razón, el examinador necesita enfocarlas con la mayor sensibilidad hacia las necesidades y derechos del individuo examinado. Para enfatizar esta cuestión crucial, se ha dedicado uno de los primeros capítulos a los elementos sutiles del proceso de prueba, incluyendo cuestiones como el establecimiento del *rapport* y la vigilancia de las influencias ambientales indeseables sobre los resultados de prueba. El último tema en el libro también enfatiza la naturaleza contractual de la evaluación al revisar las cuestiones profesionales y normas éticas en la aplicación de pruebas.

Otro tema que se destaca en este libro es la evaluación neuropsicológica, un subcampo floreciente de la psicología clínica que en la actualidad es una especialidad bien establecida con derechos propios. La evaluación neuropsicológica es, en definitiva, un área creciente y hoy en día constituye una de las principales aplicaciones de las pruebas psicológicas; se ha dedicado un capítulo completo a esta importante materia. Para que el lector pueda apreciar mejor el espectro y propósito de la evaluación neuropsicológica, se inicia el capítulo con una reseña sucinta de los principios neuropsicológicos antes de analizar los instrumentos específicos. Esta revisión introduce conceptos importantes en la evaluación neuropsicológica, como la relación entre la disfunción cerebral localizada y los síntomas conductuales específicos. Sin embargo, los lectores que necesiten dejar a un lado la sección sobre los fundamentos neurológicos de la conducta pueden hacerlo con una pérdida mínima —la sección sobre pruebas y procedimientos neuropsicológicos es comprensible de manera independiente.

Éste es más que un libro sobre pruebas y sus confiabilidades y valideces. También se exploran muchas cuestiones con gran valor que se relacionan con el conocimiento acerca de las pruebas. Las pruebas psicológicas son polémicas precisamente debido a que sus consecuencias pueden ser dañinas, con toda seguridad para los individuos y quizá también para todo el entramado social. No se han evadido las controversias que rodean su empleo. Temas independientes estudian las contribuciones genéricas y ambientales a la inteligencia, los orígenes de las diferencias raciales en CI, la parcialidad de prueba y las preocupaciones externas a la validez, el falseamiento de resultados en pruebas grupales de rendimiento, el testimonio en tribunales y las cuestiones éticas relacionadas con las pruebas psicológicas.

NOTA SOBRE LOS EJEMPLOS DE CASO

Esta edición continúa utilizando historias clínicas y viñetas que presentan conceptos sobre las pruebas e ilustran la aplicación ocasionalmente abusiva de las pruebas psicológicas. Estos ejemplos se colocan en "apartados" y se les

denomina Ejemplos de Caso. La mayor parte se basan en experiencias personales más que en trabajos académicos. Todas estas historias clínicas son reales; sucedieron en realidad —lo sé porque tengo conocimiento directo sobre la veracidad de cada anécdota. Estas cuestiones merecen que se les destaque porque es probable que al lector le parezca que algunas son francamente un producto de la fantasía y casi más allá de todo crédito. Por supuesto, a fin de garantizar el anonimato de las personas e instituciones, se han alterado ciertos detalles poco esenciales y se conserva la idea general básica de los acontecimientos originales.

CAMBIOS CON RESPECTO A LA SEGUNDA EDICIÓN EN INGLÉS

Por supuesto, una edición revisada se esforzaría en añadir las últimas conclusiones acerca de pruebas específicas y en incorporar nuevos conceptos a los anteriores. Aquí se han hecho ambas cosas. Por ejemplo, en los últimos años se han revisado muchas pruebas (WAIS-III, Leiter-R, SIB-R, para nombrar sólo algunas), y en este libro se han descrito las ediciones más recientes y reseñado la investigación relevante. También, algunos conceptos han atravesado por una reinterpretación, por lo que se han incluido las perspectivas más novedosas sobre ellos. Un caso es la prueba alfa de Cronbach, que en general se interpreta como un índice de la unidimensionalidad de una prueba o escala, pero ahora se considera como un índice de la homogeneidad de los reactivos (y no necesariamente como una medida de la unidimensionalidad de la prueba o escala).

Además de actualizar las pruebas y conceptos, la tercera edición en inglés presenta algunos cambios en cobertura de temas; los importantes son los siguientes:

1. Las diferentes evaluaciones basadas en la escuela se han agrupado en el Tema 10A: Evaluación escolar.
2. Se aumentó la cobertura sobre evaluación forense; se estudia en el Tema 10B, Aplicaciones forenses de la evaluación.
3. Por primera vez se incluye la evaluación espiritual y religiosa en el Tema 12B: Valores en el desarrollo moral, espiritual y religioso.
4. Las predicciones sobre el futuro de las pruebas se han incluido en el Tema 15A: Evaluación por computadora y el futuro de las pruebas.
5. Se aumentó el estudio acerca de las cuestiones éticas y sociales en el Tema 15B: Cuestiones éticas y sociales en las pruebas.

ESQUEMA DEL LIBRO

ORGANIZACIÓN TEMÁTICA

Para orientar al mayor público posible, se ha incorporado un esquema que divide el gigantesco campo de las pruebas psicológicas —su historia, principios y aplicaciones— en 30 temas pequeños, manejables y modulares. Es asombroso descubrir que, en general, los 30 temas se agrupan en pares; así, el lector notará que el libro también se organiza en una serie ordenada de 15 capítulos con dos temas cada uno. El formato de los capítulos ayuda a identificar los pares de temas que se relacionan, y reduce la necesidad de preámbulos redundantes para cada tema.

La unidad más fundamental e indivisible del libro es el tema. Cada uno tiene una posición independiente y el lector encuentra en ellos un número manejable de conceptos y la reseña de pruebas. Para el estudiante, la ventaja de la organización temática es que los temas individuales son suficientemente pequeños como para leerlos de una sola vez; para el instructor, la ventaja es que las materias que se consideran de menor importancia se pueden eliminar con facilidad de la lista de lecturas. Por supuesto, preferiría que todos los estudiantes leyeran todos los temas, pero hay que ser realista. Con frecuencia se requiere un texto condensado por razones prácticas, como la longitud del periodo

escolar. En esos casos, el instructor tendrá facilidad para establecer un subconjunto de temas que satisfagan las necesidades curriculares de casi cualquier curso sobre pruebas psicológicas.

ESQUEMA BÁSICO

Los 15 capítulos se dividen en seis áreas amplias, como se presenta a continuación:

Historia

Capítulo 1. Historia de las pruebas psicológicas
- Tema 1A. Orígenes de las pruebas psicológicas
- Tema 1B. Primeras pruebas en EUA

Fundamentos

Capítulo 2. Pruebas y su proceso de aplicación
- Tema 2A. Naturaleza y usos de las pruebas psicológicas
- Tema 2B. Proceso de aplicación

Capítulo 3. Normas y confiabilidad
- Tema 3A. Normas y estandarización
- Tema 3B. Conceptos de confiabilidad

Capítulo 4. Validez y desarrollo de pruebas
- Tema 4A. Conceptos básicos de validez
- Tema 4B. Construcción de pruebas

Pruebas de inteligencia y de capacidad

Capítulo 5. Pruebas de inteligencia I: Teorías y evaluación preescolar
- Tema 5A. Teorías y medición de la inteligencia
- Tema 5B. Evaluación de las capacidades de lactantes y preescolares

Capítulo 6. Pruebas de inteligencia II: Pruebas individuales y grupales
- Tema 6A. Pruebas individuales de inteligencia
- Tema 6B. Pruebas grupales de inteligencia

Capítulo 7. Sesgo de una prueba y examen de poblaciones especiales
- Tema 7A. Pruebas con poblaciones especiales
- Tema 7B. Sesgo de una prueba y otras controversias

Capítulo 8. Pruebas grupales de aptitud y de rendimiento
- Tema 8A. Pruebas de aptitud y análisis factorial
- Tema 8B. Pruebas grupales de rendimiento

Aplicaciones especializadas

Capítulo 9. Evaluación neuropsicológica y geriátrica
- Tema 9A. Compendio de neuropsicología
- Tema 9B. Evaluación neuropsicológica y geriátrica

Capítulo 10. Ambientes especiales para la evaluación psicológica
- Tema 10A. Evaluación escolar
- Tema 10B. Aplicaciones forenses de la evaluación

Capítulo 11. Evaluación industrial y organizacional
- Tema 11A. Evaluación y selección de personal
- Tema 11B. valoración del desempeño en el trabajo

Capítulo 12. Evaluación vocacional y de valores
- Tema 12A. Evaluación de intereses y valores de trabajo
- Tema 12B. Valores en el desarrollo moral, espiritual y religioso

Pruebas de personalidad

Capítulo 13. Orígenes de las pruebas de personalidad
- Tema 13A. Teorías y medición de la personalidad
- Tema 13B. Técnicas proyectivas

Capítulo 14. Evaluación estructurada de la personalidad
- Tema 14A. Inventarios de autoinforme
- Tema 14B. Evaluación conductual y enfoques relacionados

Evaluación por computadora y aspectos profesionales

Capítulo 15. Temas y cuestiones especiales en las pruebas
- Tema 15A. Evaluación por computadora y el futuro de las pruebas
- Tema 15B. Cuestiones éticas y sociales en las pruebas

El libro también presenta un extenso glosario, apéndices para localizar las pruebas y sus editoriales —destacando el Apéndice A; principales hitos en la historia de las pruebas psicológicas— y una tabla para convertir rangos percentiles a equivalentes en puntuación estándar y estandarizada. Por supuesto, para satisfacer sus necesidades personales, los lectores y los instructores del curso pueden elegir entre estos temas para satisfacer sus necesidades.

Contenido

Reconocimientos **V**

Prefacio **VII**

Capítulo 1. Historia de las pruebas psicológicas **1**

Tema 1A. Orígenes de las pruebas psicológicas 1

Tema 1B. Primeras pruebas en EUA 19

Capítulo 2. Pruebas y su proceso de aplicación 35
Tema 2A. Naturaleza y usos de las pruebas psicológicas 35
Tema 2B. Proceso de aplicación 51

Capítulo 3. Normas y confiabilidad 69
Tema 3A. Normas y estandarización 69
Tema 3B. Conceptos de confiabilidad 93

Capítulo 4. Validez y desarrollo de pruebas 117
Tema 4A. Conceptos básicos de validez 117
Tema 4B. Construcción de pruebas 143

Capítulo 5. Pruebas de inteligencia I: Teorías y evaluación preescolar 169
Tema 5A. Teorías y medición de la inteligencia 169
Tema 5B. Evaluación de las capacidades de lactantes y preescolares 191

Capítulo 6. Pruebas de inteligencia II: Pruebas individuales y grupales 213
Tema 6A. Pruebas individuales de inteligencia 213
Tema 6B. Pruebas grupales de inteligencia 243

Capítulo 7. Sesgo de una prueba y examen de poblaciones especiales 263
Tema 7A. Pruebas con poblaciones especiales 263
Tema 7B. Sesgo de una prueba y otras controversias 289

Capítulo 8. Pruebas grupales de aptitud y de rendimiento 321
Tema 8A. Pruebas de aptitud y análisis factorial 321
Tema 8B. Pruebas grupales de rendimiento 347

Capítulo 9. Evaluación neuropsicológica y geriátrica 363
Tema 9A. Compendio de neuropsicología 363
Tema 9B. Evaluación neuropsicológica y geriátrica 391

Capítulo 10. Ambientes especiales para la evaluación psicológica 419
Tema 10A. Evaluación escolar 419
Tema 10B. Aplicaciones forenses de la evaluación 451

Capítulo 11. Evaluación industrial y organizacional 477
Tema 11A. Evaluación y selección de personal 477
Tema 11B. Valoración del desempeño en el trabajo 507

Capítulo 12. Evaluación vocacional y de valores **527**
Tema 12A. Evaluación de intereses y valores de trabajo 527
Tema 12B. Valores en el desarrollo moral, espiritual y religioso 555

Capítulo 13. Orígenes de las pruebas de personalidad **575**
Tema 13A. Teorías y medición de la personalidad 575
Tema 13B. Técnicas proyectivas 593

Capítulo 14. Evaluación estructurada de la personalidad **615**
Tema 14A. Inventarios de autoinforme 615
Tema 14B. Evaluación conductual y enfoques relacionados 645

Capítulo 15. Temas y cuestiones especiales en las pruebas **663**
Tema 15A. Evaluación por computadora y el futuro de las pruebas 663
Tema 15B. Cuestiones éticas y sociales en las pruebas 685

Apéndice A. Principales hitos en la historia de las pruebas psicológicas **705**

Apéndice B. Direcciones de editoriales de pruebas **709**

Apéndice C. Principales pruebas y sus editoriales **713**

Apéndice D. Equivalentes de rangos percentiles en puntuación estándar y estandarizada en una distribución normal **719**

Glosario **723**

Referencias **739**

Índice de autores **791**

Índice temático **717**

CAPÍTULO 1

Historia de las pruebas psicológicas

TEMA 1A
Orígenes de las pruebas psicológicas

Importancia de las pruebas
Ejemplo de caso 1–1. Consecuencias de los resultados de prueba
Formas rudimentarias de las pruebas en China en el 2200 a.C.
Antecedentes psiquiátricos de las pruebas psicológicas
Era de los instrumentos de bronce de las pruebas
Cambio en los conceptos de retraso mental en el siglo XIX
Influencia de las primeras investigaciones de Binet sobre su prueba
Binet y las pruebas de procesos mentales superiores
Las escalas revisadas y el advenimiento del CI
Resumen

La historia de las pruebas psicológicas es fascinante y tiene gran relevancia para las prácticas actuales. Después de todo, las pruebas contemporáneas no surgieron de un vacío; evolucionaron lentamente a partir de una multitud de precursores que desfilaron a lo largo de los últimos 100 años. En vista de ello, este capítulo presenta una revisión de las raíces históricas de las pruebas psicológicas actuales. El tema 1A: Orígenes de las pruebas psicológicas, se centra en gran medida en los esfuerzos de los psicólogos europeos para medir la inteligencia durante la última parte del siglo XIX y la época anterior a la primera Guerra Mundial. Con frecuencia, estas primeras pruebas de inteligencia y sus sucesoras ejercieron poderosos efectos sobre los individuos examinados con ellas, de modo que el primer tema también incorpora un paréntesis breve que documenta la importancia de los resultados de las pruebas psicológicas. El tema 1B: Primeras pruebas en EUA, cataloga las numerosas pruebas desarrolladas por los psicólogos estadounidenses en la primera mitad del siglo XX.

Las pruebas psicológicas en su forma actual se originaron hace poco más de 100 años en los estudios de laboratorio sobre discriminación sensorial, habilidades motoras y tiempo de reacción. El genio británico Francis Galton (1822–1911) inventó la primera batería de pruebas, un conjunto peculiar de medidas sensoriales y motoras que se revisará más adelante. El psicólogo estadounidense James McKeen Cattell (1860–1944) estudió con Galton y después, en 1890, estipuló los temas esenciales de las pruebas modernas en su artículo clásico titulado *Mental Tests and Measurements*. Se mostró cauto y modesto al describir los propósitos y aplicaciones de sus instrumentos:

> La psicología no puede lograr la certidumbre y exactitud de las ciencias físicas, a menos que se fundamente en los experimentos y la medición. Se puede dar un paso en dicha dirección al aplicar una serie de pruebas y medidas mentales a un gran número de individuos. Los resultados tendrían un valor científico considerable en el descubrimiento de la constancia de los procesos mentales, su interdependencia y su variación bajo circunstancias diferentes. Además, los individuos encontrarían sus pruebas interesantes y, quizá, útiles con respecto al entrenamiento, modo de vida o indicación de enfermedad. El valor científico y práctico de dichas pruebas aumentaría en gran medida si se adoptara un sistema uniforme, de modo que pudieran compararse y combinarse las determinaciones realizadas en diferentes momentos y lugares (Cattell, 1890).

La conjetura de Cattell de que "quizá" las pruebas podrían ser útiles en "el entrenamiento, modo de vida o indicación de enfermedad" debe clasificarse con toda seguridad como una de las subestimaciones proféticas más notables de todos los tiempos. Cualquier persona criada en el mundo occidental sabe que las pruebas psicológicas han surgido de sus tímidos inicios para volverse un gran negocio y una institución cultural que permea a la sociedad moderna. Para citar tan sólo un ejemplo, considérese el número de pruebas estandarizadas de rendimiento y capacidad aplicadas en los sistemas escolares de EUA. Aunque es difícil obtener datos exactos en cuanto a la cantidad de dichas pruebas, es probable que no sea exagerado estimar 200 millones por año (Medina y Neill, 1990). Por supuesto, el número total de pruebas aplicadas anualmente también incluye millones de pruebas de personalidad y cantidades incontables de miles de otros tipos de pruebas que existen en la actualidad (Conoley y Kramer, 1989, 1992; Mitchell, 1985; Sweetland y Keyser, 1987). No hay duda de que las pruebas son un método generalizado, pero ¿esto causa algún impacto?

IMPORTANCIA DE LAS PRUEBAS

Las pruebas se utilizan en casi todos los países con propósitos de orientación psicológica, selección y asignación. Su aplicación ocurre en entornos tan diversos como escuelas, servicio público, industria, clínicas médicas y centros de orientación psicológica. La mayoría de las personas han tomado docenas de pruebas y no le dan gran importancia al asunto. Sin embargo, para el momento en que el individuo típico llega a la edad de retiro, es probable que los resultados de las pruebas psicológicas hayan ayudado a moldear su destino. Los cambios en el curso de la vida debidos a los resultados de pruebas psicológicas podrían ser sutiles, como ocurre cuando un futuro matemático accede a un curso de cálculo avanzado con base en calificaciones de rendimiento del primer año de preparatoria. Es más común que los resultados de pruebas psicológicas alteren el destino individual de manera profunda. El que se acepte a una persona en una universidad y no en otra; que se le ofrezca un empleo, pero se le rechace en otro; se le diagnostique como deprimido o no, todas estas determinaciones dependen, al menos en parte, de la interpretación de los resultados de pruebas que realizan individuos con autoridad. Dicho en términos sencillos, los resultados de las pruebas psicológicas cambian la vida. Por tal razón es prudente —y, de hecho, casi obligatorio— que los estudiantes de psicología aprendan acerca de los usos

EJEMPLO DE CASO 1–1
CONSECUENCIAS DE LOS RESULTADOS DE PRUEBA

La importancia de las pruebas psicológicas se ilustra mejor con un ejemplo. Considérense las siguientes viñetas breves:

- Una niña tímida y retraída de siete años de edad recibe una prueba de CI que le aplica un psicólogo educativo. Su puntuación es mucho más alta de lo que el maestro esperaba. Se admite a la alumna en un programa para niños dotados y talentosos donde florece hasta convertirse en una estudiante confiada en sí misma y gregaria.
- Tres niños de una familia que vive cerca de una fundición de plomo se ven expuestos a los efectos tóxicos del polvo de plomo y sufren daño neurológico. En parte, con base en los resultados de una prueba psicológica que demuestran alteración en la inteligencia y una breve capacidad de atención en los niños, la familia recibe una compensación de 8 millones de dólares (EUA) de parte de la compañía que es propietaria de la fundición.
- A un candidato al puesto de oficial de policía se le aplica un inventario de personalidad como parte del proceso de selección. La prueba indica que el candidato tiende a actuar sin pensar y resiste la supervisión de parte de figuras de autoridad. Aun cuando tiene un entrenamiento excelente e impresiona a los entrevistadores, el candidato no recibe la oferta de trabajo.
- Una estudiante, insegura de cuál carrera debe seguir, responde un inventario de intereses vocacionales. La prueba indica que debería trabajar como farmacéutica. Se inscribe en un curso propedéutico de química farmacéutica, pero encuentra que las clases son tanto difíciles como aburridas. Después de tres años abandona esta carrera para tomar una de danza, frustrada por el hecho de que todavía enfrenta tres años adicionales en la universidad para obtener el título.
- Un solicitante a una escuela de posgrado en psicología clínica contesta el Inventario Multifásico de Personalidad (MMPI). Sus recomendaciones y promedio de calificaciones son notables, sin embargo, debe pasar el último obstáculo que representa el MMPI. Sus resultados son razonablemente normales, pero un tanto defensivos; en una estrecha votación, el comité de aceptación le extiende una invitación. Irónicamente, ésta es la única escuela de posgrado que lo admite —otras 19 lo rechazan. Acepta la invitación y se fascina con el estudio de la evaluación psicológica. Muchos años después, escribe este libro.

actuales y de los abusos ocasionales en la aplicación de pruebas. En el ejemplo de caso 1–1 se ilustran los cambios de vida que son consecuencia de las pruebas psicológicas a través de varias muestras de historias clínicas verdaderas.

La importancia de las pruebas también es evidente desde la perspectiva de una revisión histórica. En general, los estudiantes de psicología consideran los temas históricos como aburridos, áridos y difíciles y, en ocasiones, dichos prejuicios están bien justificados. Después de todo, muchos libros de texto no logran explicar la relevancia de las cuestiones históricas y proporcionan sólo bocetos vagos de los desarrollos iniciales en las pruebas mentales. Como resultado, es frecuente que los alumnos de psicología concluyan de manera incorrecta que los temas históricos son aburridos e irrelevantes.

En realidad, la historia de las pruebas psicológicas es fascinante y tiene relevancia sustancial para las prácticas actuales. La evolución

histórica es pertinente para las pruebas contemporáneas debido a las siguientes razones:

1. Una revisión de los orígenes de las pruebas psicológicas ayuda a explicar prácticas actuales que, de otra manera, podrían parecer arbitrarias e incluso peculiares. Por ejemplo, ¿por qué muchas pruebas de inteligencia actuales incorporan una capacidad en apariencia no intelectual como la memoria a corto plazo de dígitos? La respuesta es, en parte, la inercia histórica —las pruebas de inteligencia siempre han incluido una medida de retención de dígitos.
2. El poder y las limitaciones de las pruebas también resaltan con mayor facilidad cuando dichos métodos se observan dentro de un contexto histórico. Por ejemplo, el lector descubrirá que las pruebas de inteligencia modernas son excepcionalmente buenas para pronosticar el fracaso en la escuela, debido precisamente a que éste fue el propósito original y único del primero de dichos instrumentos desarrollado en París, Francia, a principios de este siglo.
3. Por último, la historia de las pruebas psicológicas contiene algunos episodios tristes y lamentables que ayudan a recordar que no se debe ser demasiado vehemente en el uso que en la actualidad se da a las pruebas. Por ejemplo, con base en la aplicación insensata y prejuiciada de los resultados de pruebas de inteligencia, varios psicólogos destacados ayudaron a aprobar el Decreto de 1924 de Restricción a la Inmigración *(Immigration Restriction Act of 1924).*

En capítulos posteriores se analizarán los principios de las pruebas psicológicas, se investigarán las aplicaciones a campos específicos (p. ej., personalidad, inteligencia, neuropsicología) y se reflexionará acerca de las consecuencias sociales y legales de las pruebas. Sin embargo, el lector encontrará que estos temas son más comprensibles cuando se les analiza en un contexto histórico. Así, por el momento, se comenzará por la revisión de las formas rudimentarias de prueba que existieron hace más de 4 000 años en la China imperial.

FORMAS RUDIMENTARIAS DE LAS PRUEBAS EN CHINA EN EL AÑO 2200 a.C.

Aunque el empleo extendido de las pruebas psicológicas es, en gran medida, un fenómeno del siglo XX, los historiadores señalan que las formas rudimentarias de las pruebas datan de, cuando menos, el año 2200 a.C., cuando el emperador chino hizo que se examinara a sus oficiales de gobierno cada tercer año para determinar su idoneidad para el puesto (Chaffee, 1985; DuBois, 1970; Franke, 1963; Lai, 1970; Teng, 1942–43). Dichas pruebas se modificaron y perfeccionaron a lo largo de siglos hasta que se introdujeron exámenes por escrito durante la dinastía Han (202 a.C.–200 d.C.). Se sometía a prueba en cinco temas: ley civil, asuntos militares, agricultura, contribuciones y geografía.

El sistema de examen chino adquirió su forma final aproximadamente en 1370, cuando se enfatizó la destreza en los clásicos escritos de Confucio. En el examen preliminar se pedía a los candidatos que pasaran un día y una noche en una pequeña cabina aislada, componiendo ensayos sobre temas asignados y escribiendo un poema. De 1 a 7% que aprobaban continuaban con los exámenes por distrito, que requerían tres sesiones independientes de tres días con sus noches.

Como es obvio, estos exámenes eran extenuantes y rigurosos, pero éste no era el último nivel. De 1 a 10% de los individuos que aprobaban tenían el privilegio de ir a Pekín para la ronda final de exámenes. Quizá 3% de este último grupo aprobaba y lograba el puesto de mandarín, elegible para el servicio público

Aunque los chinos desarrollaron la parafernalia para un programa amplio de exámenes del servicio público, las semejanzas entre sus tradiciones y las prácticas actuales de prueba son superficiales en cuanto a lo esencial. Sus prácticas de prueba no sólo eran innecesariamente agotadoras, sino que los chinos tampoco validaron sus procedimientos de selección. Sin embargo, parece ser que el programa de

exámenes incorporaba criterios pertinentes para la selección. Por ejemplo, en los exámenes escritos, la belleza de la caligrafía tenía un peso muy importante. Dadas las características sumamente estilísticas de las formas de escritura china, sin duda la buena caligrafía era esencial para la comunicación clara y exacta. Así, es probable que la caligrafía fuera un mecanismo pertinente de predicción para el empleo en el servicio público. En respuesta al amplio descontento, el sistema de examen se abolió por decreto real en 1906 (Franke, 1963).

ANTECEDENTES PSIQUIÁTRICOS DE LAS PRUEBAS PSICOLÓGICAS

La mayoría de los historiadores ubican el comienzo de las pruebas psicológicas en la investigación experimental sobre diferencias individuales que floreció en Alemania y Gran Bretaña a finales del siglo XIX. No hay duda de que los primeros experimentalistas como Wilhelm Wundt, Francis Galton y James McKeen Cattell establecieron los fundamentos para las pruebas del siglo XIX y se revisarán en detalle algunas de sus contribuciones. Pero las pruebas psicológicas le deben tanto a la primera psiquiatría como a los laboratorios de psicología experimental. De hecho, el examen de los enfermos mentales aproximadamente a mediados del siglo XIX resultó en el desarrollo de muchas de las primeras pruebas (Bondy, 1974), las cuales mostraban la ausencia de estandarización y, en consecuencia, se relegaron al olvido. Sin embargo, influyeron en la determinación del curso de las pruebas psicológicas, de modo que es importante mencionar algunos avances típicos de esta época.

En 1885, el médico alemán Hubert von Grashey desarrolló el antecedente del cilindro mnemónico como un medio para examinar a pacientes con lesión cerebral. Se les mostraban a los sujetos palabras, símbolos o ilustraciones a través de una ranura en una hoja de papel que se movía lentamente sobre los estímulos.Grashey encontró que muchos pacientes podían reconocer los estímulos completos, pero no identificarlos cuando se mostraban a través de la ranura en movimiento. Poco después, el psiquiatra alemán Conrad Rieger desarrolló una batería de pruebas excesivamente ambiciosa para la detección del daño cerebral, la cual requería más de 100 horas para su aplicación, por lo que pronto cayó su aceptación.

En resumen, la psiquiatría inicial contribuyó al movimiento de las pruebas mentales al mostrar que los procedimientos estandarizados pueden ayudar a revelar la naturaleza y grado de los síntoma en pacientes con enfermedades mentales y con lesión cerebral. La mayoría de las primeras pruebas desarrolladas por los psiquiatras se quedaron en el olvido, pero se estandarizaron unoscuantos procedimientos y se les conserva en variaciones modernas (Bondy, 1974).

ERA DE LOS INSTRUMENTOS DE BRONCE DE LAS PRUEBAS

La psicología experimental floreció a finales del siglo XIX en la Europa continental y Gran Bretaña. Por primera vez en la historia, los psicólogos se alejaron de los métodos totalmente subjetivos e introspectivos que se habían seguido de manera tan infructuosa en siglos anteriores. En lugar de ello, las capacidades humanas se sometieron a prueba en laboratorios. Los investigadores utilizaron procedimientos objetivos que tenían posibilidad de replicarse. Lejos quedaron los días en que laboratorios rivales tenían grandes discusiones acerca del "pensamiento sin imágenes", donde un grupo afirmaba que existía, mientras otro decía que tal acontecimiento mental era imposible.

Aunque el nuevo énfasis en los métodos objetivos y las cantidades medibles era una gran mejoría sobre el mentalismo generalmente estéril que le precedió, la nueva psicología experimental era, en sí misma, un callejón sin salida, al menos en cuanto a pruebas psicológicas se refiere. El problema estaba en que los primeros psicólogos experimentales confundieron los

simples procesos sensoriales con la inteligencia. Utilizaron diversos instrumentos de bronce para medir umbrales sensoriales y tiempos de reacción, pensando que dichas capacidades eran esenciales para la inteligencia. De aquí que en ocasiones se llame a este periodo la era de los Instrumentos de Bronce de las pruebas psicológicas.

A pesar del falso comienzo de los primeros experimentalistas, al menos proporcionaron una metodología apropiada a la psicología. Pioneros como Wundt, Galton, Cattell y Wissler mostraron que era posible exponer a la mente al escrutinio y la medición científicos. Éste fue un cambio decisivo en las suposiciones axiomáticas de la psicología, un cambio que ha permanecido hasta la actualidad.

La mayor parte de las fuentes le atribuyen a Wilhelm Wundt (1832–1920) la fundación del primer laboratorio psicológico en 1879 en Leipzig, Alemania. Es menos sabido que este investigador llevaba años midiendo procesos mentales, al menos desde 1862, cuando experimentó con su **medidor del pensamiento** (Diamond, 1980). Este instrumento era un péndulo calibrado con agujas que sobresalían a cada lado. El péndulo oscilaba de un lado a otro, golpeando campanas con las agujas. La tarea del observador consistía en tomar nota de la posición del péndulo cuando las campanas sonaban. Por supuesto, Wundt podía ajustar de antemano las agujas y, por tanto, conocer la posición precisa del péndulo cuando tocaba cada campana. Wundt consideraba que la diferencia entre la posición observada del péndulo y la posición real podía proporcionar un medio para determinar la velocidad de pensamiento del observador.

El análisis de Wundt se relacionaba con un añejo problema en astronomía, que consistía en que dos o más astrónomos que utilizaran de manera simultánea el mismo telescopio (con múltiples oculares) informarían diferentes tiempos de travesía, a medida que las estrellas atravesaban una línea de rejilla en el telescopio. Incluso en tiempos de Wundt, un acontecimiento muy conocido en la historia de la ciencia era que Kinnebrook, un ayudante en el Observatorio Real en Inglaterra, había sido despedido en 1796 debido a que sus tiempos de travesía estelar estaban demorados en casi un segundo (Boring, 1950). El análisis de Wundt ofreció otra explicación que no suponía la incompetencia por parte de nadie. Dicho en términos sencillos, consideraba que la velocidad de pensamiento podía diferir de una persona a otra:

> Para cada persona debe haber una cierta velocidad de pensamiento, que nunca puede exceder debido a su constitución mental dada. Así, de la misma manera en que una máquina de vapor puede ir más rápido que otra, es probable que esta velocidad de pensamiento no sea la misma en todas las personas (Wundt, 1862, según la traducción de Rieber, 1980).

Este análisis de los informes de tiempos en telescopio parece simplista según normas actuales, y pasa por alto la posible contribución de factores como la atención, motivación y retroalimentación autocorrectora de los ensayos anteriores. En su lado positivo, éste fue al menos un análisis empírico que buscaba explicar las diferencias individuales, en vez de tratar de deshacerse de ellas por medio de explicaciones. Y ésa es la relevancia para las prácticas actuales en cuanto a pruebas psicológicas. A pesar de lo rudimentario, Wundt midió los procesos mentales y reconoció a regañadientes las diferencias individuales.*

GALTON Y LA PRIMERA BATERÍA DE PRUEBAS MENTALES

Sir Francis Galton (1822–1911) fue el pionero de la nueva psicología experimental en la Gran Bretaña del siglo XIX. Estaba obsesionado con la medición y su carrera intelectual parece haberse visto dominada por la creencia de que podía medirse casi cualquier cosa. Sus intentos por medir el intelecto a través del tiempo de

* Este énfais en las diferencias individuales era poco común para Wundt. Él es más reconocido por proponer leyes generales del pensamiento para la mente adulta promedio.

reacción y de tareas de discriminación sensorial son muy conocidos. Sin embargo, para apreciar sus amplios intereses, el lector debería estar enterado de que Galton también diseñó técnicas para medir la belleza, la personalidad, el nivel de aburrimiento de una lectura y la eficacia de la oración, por nombrar sólo unas cuantas de las empresas que su biógrafo ha catalogado con gran detalle (Pearson, 1914, 1924, 1930ab).

Galton fue un genio que estaba más interesado en los problemas de la evolución humana que en la psicología en sí (Boring, 1950). Sus dos obras con mayor influencia fueron *Hereditary Genius* (1869), un análisis empírico que tenía el propósito de probar que los factores genéticos son abrumadoramente importantes para el logro de la superioridad, e *Inquiries into Human Faculty and its Development* (1883), una serie desigual de ensayos que enfatizaban las diferencias individuales en cuanto a facultades mentales.

Boring (1950) considera que *Inquiries* es el comienzo del movimiento de las pruebas mentales y del advenimiento de la psicología científica de las diferencias individuales. El libro es una curiosa mezcla de investigación empírica y ensayos especulativos acerca de temas tan diversos como la "diferencia apenas perceptible" en el aumento de peso y la disminución de la fertilidad entre animales que son producto de endogamia. Sin embargo, existe un tema común que une a estos ensayos diversos; Galton demostró una y otra vez que las diferencias individuales no sólo existen, sino que se pueden medir de manera objetiva.

Galton tomó prestados los lentos procedimientos psicofisicos practicados por Wundt y otros investigadores en el continente europeo, y los adaptó a una serie de medidas sensoriomotoras simples y rápidas. Así, continuó con la tradición de los instrumentos de bronce en las pruebas mentales, pero con una importante diferencia: sus procedimientos fueron más adecuados para la oportuna recolección de datos de cientos, si no es que miles, de sujetos. Debido a sus esfuerzos para diseñar medidas viables de las diferencias individuales, en general, los historiadores de las pruebas psicológicas consideran a Galton como el padre de las pruebas mentales (Goodenough, 1949; Boring, 1950).

Para proseguir con su estudio de las diferencias individuales, Galton estableció un laboratorio psicométrico en Londres en la *International Health Exhibition* de 1884. Después se le transfirió al *London Museum*, donde se mantuvo durante seis años. Varias medidas antropométricas y psicométricas estaban distribuidas en una larga mesa, a un lado de una estrecha habitación. Los participantes ingresaban por un extremo, a tres peniques por entrada, y se les aplicaban pruebas sucesivas a medida que pasaban junto a la mesa. Se sometió a prueba al menos a 17 000 individuos durante los decenios de 1880 y 1890. Aproximadamente 7 500 de los registros de datos individuales han sobrevivido hasta la actualidad (Johnson, McClearn, Yuen, Nagoshi, Ahern y Cole, 1985).

Las pruebas y medidas implicaban tanto el campo físico como el conductual. Las características físicas evaluadas eran estatura, peso, largo de la cabeza, ancho de la misma, largo del brazo, largo del dedo cordial y del antebrazo, entre otras. Las pruebas conductuales incluían fortaleza de la prensión ejercida con la mano determinada con un dinamómetro; capacidad vital de los pulmones medida con un espirómetro, agudeza visual; tono más alto audible; velocidad de soplo y tiempo de reacción (TR) tanto para estímulos visuales como auditivos.

En última instancia, los intentos simplistas de Galton por estimar el intelecto con medidas de tiempo de reacción y discriminación sensorial resultaron infructuosos. Sin embargo, proporcionó un enorme ímpetu al movimiento de evaluación al demostrar que podían diseñarse pruebas objetivas y obtenerse puntuaciones significativas a través de procedimientos estandarizados.

CATTELL IMPORTA LOS INSTRUMENTOS DE BRONCE A EUA

James McKeen Cattell (1860-1944) estudió la nueva psicología experimental tanto con Wundt como con Galton, antes de establecerse en la Columbia University, donde, durante 26 años,

fue el decano indiscutible de la psicología estadounidense. Con Wundt realizó una serie de estudios sobre TR elaborados de manera concienzuda (1880-1882), en los que se midieron con gran precisión las fracciones de segundo que supuestamente se requerían para las diferentes reacciones mentales. También observó, casi de modo incidental, que él y otro colega tenían diferencias pequeñas pero consistentes en TR. Cattell le propuso a Wundt que dichas diferencias individuales deberían estudiarse de manera sistemática. Aunque Wundt reconoció las diferencias individuales, estaba más inclinado filosóficamente hacia el estudio de las características generales de la mente y no le prestó apoyo a la propuesta de Cattell (Fancher, 1985).

Pero Cattell recibió ayuda entusiasta para su estudio de las diferencias individuales de parte de Galton, quien apenas había inaugurado su laboratorio psicométrico en Londres. Después de tener correspondencia con Galton durante algunos años, Cattell obtuvo una beca de dos años en Cambridge, de modo que pudiera continuar con el estudio de las diferencias individuales. Cattell abrió su propio laboratorio de investigación y desarrolló una serie de pruebas que eran principalmente extensiones y adiciones a la batería de Galton.

Cattell (1890) inventó el término **prueba mental** en su famoso artículo intitulado *"Mental Tests and Measurements"*. Este artículo describía su programa de investigación, detallando 10 pruebas mentales que él proponía para su utilización con el público en general. Estas pruebas eran con toda claridad reelaboraciones perfeccionadas de la tradición de Galton:

- Fortaleza de la prensión ejercida con la mano, medida con un dinamómetro.
- Tasa de movimiento de la mano a través de una distancia de 50 cm.
- Umbral de tacto de dos puntos —distancia mínima en que dos puntos todavía se perciben como separados.
- Grado de presión necesaria para causar dolor —presión de una punta de goma contra la frente.
- Diferenciación de peso —discernir los pesos relativos de cajas de apariencia idéntica que variaban en 1 g desde los 100 hasta los 110.
- Tiempo de reacción para sonido —utilizando un instrumento similar al de Galton.
- Tiempo para nombrar colores.
- División en dos partes de una línea de 50 cm.
- Juicio de 10 seg de tiempo.
- Número de letras repetidas en una audición.

La fortaleza de la prensión ejercida con la mano parece una adición curiosa a una batería de pruebas mentales, un punto al que Cattell (1890) le brindó atención directa en su artículo. Tenía la opinión de que era imposible separar la energía corporal de la mental. Así, desde la perspectiva de Cattell, una medida ostensiblemente fisiológica como la prensión en el dinamómetro, también era un índice del propio poder mental. Es claro que la predisposición fisiológica y sensorial de toda la batería de pruebas refleja su fuerte herencia galtoniana (Fancher, 1985).

En 1891, Cattell aceptó un puesto en la *Columbia University*, que en aquel entonces era la universidad más grande de EUA. Su posterior influencia sobre la psicología estadounidense superó en mucho a sus resultados científicos individuales y se expresó en gran parte a través de sus numerosos e influyentes alumnos (Boring, 1950). Entre sus muchos y famosos estudiantes de doctorado y los años en que obtuvieron sus títulos se encuentran E. L. Thorndike (1898), quien realizó contribuciones monumentales a la teoría del aprendizaje y a la psicología educativa; R. S. Woodworth (1899), quien fue el autor del muy popular e influyente *Experimental Psychology* (1938); y E. K. Strong (1911), cuyo *Vocational Interest Blank* —revisado desde entonces— todavía tiene amplio uso. Pero entre los alumnos de Cattell, probablemente fue Clark Wissler (1901) quien tuvo la mayor influencia en la historia inicial de las pruebas psicológicas.

Wissler obtuvo tanto las puntuaciones en prueba mental como las calificaciones académicas de más de 300 estudiantes de la *Columbia*

University y el *Barnard College*. Su meta consistía en demostrar que los resultados de prueba podían pronosticar el desempeño académico. Con la perspectiva de finales del siglo XX acerca de la investigación y las pruebas, parece sorprendente que los primeros experimentalistas esperaran tanto tiempo para realizar dicha investigación básica en cuanto a validación. Los resultados de Wissler (1901) mostraron que virtualmente no había una tendencia a que las puntuaciones de una prueba mental se correlacionaran con el rendimiento académico. Por ejemplo, la posición dentro de la clase se correlacionó en .16 con la memoria de listas de números, –.08 con la fortaleza en el dinamómetro, .02 con el nombrado de colores y –.02 con el tiempo de reacción. La correlación más alta (.16) fue estadísticamente significativa debido al gran tamaño de la muestra. Sin embargo, una correlación tan pequeña lleva consigo muy poca utilidad predictiva.*

También fueron dañinas para el movimiento de pruebas con instrumentos de bronce las insuficientes correlaciones entre las pruebas mentales en sí. Por ejemplo, la indicación de colores y la velocidad del movimiento de la mano se correlacionaron sólo en .19, mientras el TR y la indicación de colores se correlacionaron en –.15. Algo que no produce sorpresa es que varias medidas físicas, como el tamaño de la cabeza (una medida que se había conservado desde el tiempo de Galton), tampoco se correlacionaron con diversas medidas sensoriales y de TR.

Con la publicación de los desalentadores resultados de Wissler (1901), los psicólogos experimentales abandonaron, en términos generales, el uso del TR y de la discriminación sensorial como medidas de la inteligencia. Desde un punto de vista, este alejamiento del enfoque de los instrumentos de bronce fue una evolución deseable en la historia de las pruebas psicológicas. Con ello se facilitó el camino para la aceptación inmediata de las medidas más sensatas y útiles de Binet para los procesos mentales superiores.

Pero en otros sentidos, el abandono del TR y de las medidas sensoriales fue prematuro y desafortunado. Después de todo, según normas actuales, los métodos de investigación de Wissler revelaron una inocencia psicométrica extraordinaria. Al utilizar como sujetos sólo a alumnos universitarios brillantes, Wissler había introducido de manera inadvertida una restricción extrema del rango, que reduciría invariablemente el tamaño de sus correlaciones. Si se hubiera utilizado una muestra más heterogénea de sujetos, las correlaciones hubieran sido sustancialmente más grandes. Además, ciertas medidas, como el TR, eran poco confiables a nivel inherente debido al pequeño número de ensayos por sujeto. Tal falta de confiabilidad en una medida también causa graves restricciones a los límites superiores de los coeficientes de correlación.

Si los hallazgos negativos de Wissler (1901) se hubieran analizado con mayor escepticismo, no hubieran pasado 70 años para que el TR resurgiera como una medida intelectual potencialmente útil. Las correlaciones de –.40 entre las formas complejas de TR y la inteligencia no son poco comunes en absoluto (Jensen, 1982).†

Pero eso sería adelantarse en la historia. La reacción más común entre los psicólogos a inicios del siglo XX fue la de concluir a regañadientes que Galton había estado equivocado en sus intentos por inferir capacidades complejas, a partir de capacidades simples. Goodenough (1949) ha vinculado el enfoque de Galton con

* El coeficiente de correlación se analiza con mayor detalle en el tema 3B, Conceptos de Confiabilidad. Pero con el propósito de un avance rápido, las correlaciones pueden ir de –1.0 a +1.0. Los valores cercanos a 0 indican una relación lineal débil, despreciable, entre las dos variables. Por ejemplo, las correlaciones entre –.20 y +.20 se consideran, en general, de un valor mínimo para propósitos de predicción individual. Obsérvese también que las correlaciones negativas indican relación inversa.

† Las correlaciones son negativas debido a que las bajas puntuaciones en TR se relacionan con altas puntuaciones en pruebas de inteligencia.

"inferir la naturaleza del genio a partir de la naturaleza de la estupidez o las cualidades del agua a partir de las del hidrógeno y el oxígeno de los que se compone". En apariencia, los psicólogos académicos estuvieron de acuerdo con dicha autora y los intentos estadounidenses por desarrollar pruebas de inteligencia cesaron virtualmente a principios de este siglo. Por su parte, aparentemente Wissler se desalentó de tal modo por sus propios resultados que de inmediato cambió a la antropología, donde se convirtió en un fuerte ambientalista en la explicación de las diferencias entre grupos étnicos.

El vacío creado por el abandono de la tradición de Galton no duró mucho tiempo. En Europa, Alfred Binet estaba a punto de lograr un gran avance en el examen de la inteligencia. Binet introdujo su escala de inteligencia en 1905 y poco después H. H. Goddard la importó a EUA, donde se aplicó de una manera que Gould (1981) ha descrito como "el desmantelamiento en EUA de las intenciones de Binet". La duda relativa a si los psicólogos estadounidenses de principios del siglo XX trastocaron las intenciones de Binet, constituye una pregunta importante que se revisará en el siguiente tema. Pero primero se examinarán los cambios sociales en la Europa del siglo XIX que crearon la necesidad de pruebas prácticas de la inteligencia.

CAMBIO EN LOS CONCEPTOS DE RETRASO MENTAL EN EL SIGLO XIX

Muchos de los grandes inventos se han desarrollado en respuesta a las necesidades prácticas creadas por los cambios en los valores sociales. Tal es el caso de las pruebas de inteligencia. Para ser específicos, Binet desarrolló la primera de dichas pruebas a comienzos del siglo XX, para ayudar a identificar a los niños dentro del sistema escolar de París que no estaban en probabilidad de beneficiarse de la educación habitual. Antes de ese momento había poco interés en las necesidades educativas de los niños con retraso. Así, un nuevo humanismo hacia estos individuos creó el problema práctico —identificar a aquellas personas con necesidades especiales— que las pruebas de Binet habrían de resolver.

El mundo occidental a finales del siglo XIX surgía apenas de siglos de indiferencia y hostilidad hacia aquellos que tenían impedimentos psiquiátricos y mentales. Los médicos empezaban a reconocer una distinción entre personas con discapacidad emocional y con retraso mental. Por siglos, todos estos parias sociales recibían un tratamiento similar. En la edad media se les "diagnosticaba" en ocasiones como brujos y se les quemaba en la hoguera. Posteriormente, por épocas se les ignoraba, perseguía o torturaba. En su amplia historia sobre la psicoterapia y el psicoanálisis, Bromberg (1959) tiene un capítulo particularmente explícito sobre las diversas formas de maltrato hacia las personas con discapacidades mentales y emocionales, del cual sólo se proporcionará aquí un ejemplo. En 1698, un notable médico escribió un horripilante libro, *Flagellum Salutis*, en el que se abogaba por las golpizas para el tratamiento de "la melancolía, el frenesí, la parálisis, la epilepsia, la expresión facial de imbecilidad" (Bromberg, 1959).

A principios del siglo XIX, las mentes más sanas comenzaron a prevalecer. Los médicos se percataron de que algunos de aquellos individuos con alteraciones psiquiátricas tenían enfermedades reversibles que no necesariamente implicaban un intelecto disminuido, mientras otras personas excepcionales, aquéllos con retraso mental, mostraban mayor continuidad relacionada con el desarrollo e invariablemente tenían deterioro intelectual. Además, un humanismo reciente comenzó a influir en las prácticas sociales hacia los individuos con impedimentos psicológicos y mentales, con lo que surgió mayor interés en el diagnóstico y tratamiento del retraso mental. Al frente de estos desarrollos se encontraban dos médicos franceses, J. E. D. Esquirol y O. E. Sequin, cada uno de los cuales revolucionó el pensamiento acerca del retraso mental, lo que ayudó a crear la necesidad para las pruebas de Binet.

ESQUIROL Y EL DIAGNÓSTICO DEL RETRASO MENTAL

Cerca de los inicios del siglo XIX, muchos médicos habían comenzado a percibir la diferencia entre retraso mental (llamado entonces **idiotez**) y enfermedad mental (que con frecuencia se llamaba **demencia**). J. E. D. Esquirol (1772–1840) fue el primero en establecer por escrito la diferencia. Su avance diagnóstico consistió en señalar que el retraso mental era un fenómeno del desarrollo que duraba toda la vida, mientras la enfermedad mental tenía por lo general un brote más abrupto en la edad adulta. Consideraba que el retraso mental era incurable, mientras la enfermedad mental podría mostrar mejoría (Esquirol, 1845/1838).

Esquirol le dio gran énfasis a las habilidades del lenguaje en el diagnóstico del retraso mental. Esto puede ofrecer una explicación parcial de la razón por la que las posteriores pruebas de Binet y sus descendientes de tiempos modernos brindan un peso tan enorme a las capacidades lingüísticas. Después de todo, el uso original de las escalas de Binet era, primordialmente, para identificar a niños con retraso mental que no estarían en probabilidad de beneficiarse de la instrucción escolar habitual.

Esquirol también propuso el primer sistema de clasificación para el retraso mental y no debería causar sorpresa que las habilidades del lenguaje fueran los principales criterios diagnósticos. Este autor reconoció tres niveles de retraso mental: 1) aquellos que utilizan frases cortas; 2) los que utilizan sólo monosílabos y 3) aquellos que sólo lloran y no hablan. En apariencia, Esquirol no reconoció lo que ahora se llamaría **retraso mental leve** y en lugar de ello proporcionó criterios para los equivalentes de las clasificaciones actuales de retraso mental moderado, grave y profundo.

SEGUIN Y LA EDUCACIÓN DE LOS INDIVIDUOS CON RETRASO MENTAL

Quizá más que ningún otro pionero en el campo del retraso mental, O. Edouard Seguin (1812–1880) ayudó a establecer a finales del siglo XIX un nuevo humanismo hacia las personas con retraso. Había sido alumno de Esquirol y estudiado con J. M. G. Itard (1774-1838), quien es muy conocido por su intento de entrenar durante cinco años al niño salvaje de Aveyron, un niño montaraz que había vivido en el bosque durante sus primeros 11 o 12 años (Itard, 1932/1801).

Seguin pidió prestadas las técnicas utilizadas por Itard y dedicó su vida a desarrollar programas educativos para las personas con retraso. Desde 1838 había establecido una clase experimental para individuos con retraso mental. Sus esfuerzos de tratamiento le dieron reconocimiento internacional y en un momento dado se trasladó a EUA para continuar su trabajo. En 1866 publicó *Idiocy, and Its Treatment by the Physiological Method*, el primer texto de importancia acerca del tratamiento del retraso mental. Este libro proponía un enfoque sorprendentemente moderno hacia la educación de las personas con retraso mental e incluso tocaba aspectos de lo que ahora se llamaría **modificación conductual**.

Tales fueron los antecedentes sociales e históricos que permitieron el florecimiento de las pruebas de inteligencia. Ahora la atención se vuelve hacia la invención de la prueba moderna de inteligencia por parte de Alfred Binet. Se comenzará con un análisis de las primeras influencias que moldearon su famosa prueba.

INFLUENCIA DE LAS PRIMERAS INVESTIGACIONES DE BINET SOBRE SU PRUEBA

Como sabe casi cualquier estudiante de psicología, Alfred Binet (1857–1911) inventó la primera prueba moderna de inteligencia en 1905. Lo que es menos conocido, pero de igual importancia para aquellos que pretenden comprender sus contribuciones a la psicología moderna, es que Binet fue un investigador y autor prolífico mucho antes de volver su atención hacia el examen de la inteligencia. El carácter de sus

primeras investigaciones tiene injerencia para la forma subsecuente de su muy conocida prueba de inteligencia. Para quienes buscan una amplia comprensión de su esmerada influencia, es obligatorio hacer una mención breve sobre su carrera anterior. Para mayor detalle, el lector puede consultar a DuBois (1970), Fancher (1985), Goodenough (1949), Gould (1981) y Wolf (1973).

Binet comenzó su carrera en medicina, pero se vio forzado a abandonarla debido a una crisis emocional generalizada. Cambió a psicología, donde estudió el umbral de dos puntos y mostró un ligero interés en la psicología asociacionista de John Stuart Mill (1806–1873). Después seleccionó trabajar como aprendiz con el neurólogo J. M. Charcot (1825–1893) en el famoso *Salpetriere Hospital*. Así, durante un breve periodo, la ruta profesional de Binet fue paralela a la de Sigmund Freud, quien también estudió la histeria bajo la supervisión de Charcot. En el *Salpetriere Hospital*, Binet fue coautor (junto con C. Fere) de cuatro estudios que supuestamente demostraban que la inversión de la polaridad de un imán podía inducir completos cambios en el estado de ánimo (de felicidad a tristeza) o transferir una parálisis histérica (p. ej., del lado izquierdo al derecho) en un mismo individuo bajo hipnosis. En respuesta a la crítica pública de otros psicólogos, Binet publicó después una rectificación de sus descubrimientos. Éste fue un doloroso episodio para Binet, lo cual desvió su carrera de manera temporal. Sin embargo, aprendió dos cosas a través de su vergüenza. Primero, nunca utilizó de nuevo procedimientos experimentales descuidados que dieran lugar a que las sugerencias no intencionales influyeran sobre sus resultados. Segundo, se volvió escéptico acerca del *Zeitgeist* (espíritu de los tiempos) en la psicología experimental. Aplicó ambas lecciones cuando posteriormente desarrolló sus escalas de inteligencia.

En 1891, Binet fue a trabajar a la Sorbona como asistente, sin remuneración, y comenzó una serie de estudios y publicaciones que habrían de definir su nueva "psicología individual" y que culminarían en última instancia en sus pruebas de inteligencia. Binet era un experimentalista ferviente, que con frecuencia utilizaba a sus dos hijas para examinar sus pruebas de inteligencia existentes o nuevas. Al principio tuvo un acercamiento al enfoque de Cattell sobre el examen de la inteligencia, utilizando con sus dos hijas las medidas estándar de tiempo de reacción y agudeza sensorial. Los resultados fueron incómodamente inconsistentes y difíciles de interpretar. Como podría esperarse, encontró que los tiempos de reacción de sus hijas eran, en promedio, mucho más lentos que para los adultos. A partir de estos descubrimientos, Binet concluyó que la atención era un componente clave de la inteligencia, que era en sí misma una entidad sumamente multifacética. Lo que es más, se fue decepcionando del enfoque de los instrumentos de bronce para la medición de la inteligencia, lo cual es probable que explique su utilización subsecuente de medidas de los procesos mentales superiores.

Además, los experimentos sensoperceptuales de Binet con sus hijas influyeron en gran medida sobre sus perspectivas acerca de los procedimientos adecuados de prueba:

> El experimentador está obligado, hasta cierto punto, a adaptar su método al sujeto a quien se dirige. Se deben seguir ciertas reglas cuando uno experimenta con un niño, de la misma manera que existen ciertas reglas para los adultos, las personas con histeria y los enfermos mentales. Estas reglas no están escritas en ningún lado; cada quien las aprende por sí solo y se ve remunerado en gran medida. Al cometer un error y después explicar su causa, se aprende a no cometer el error en una segunda ocasión. Con respecto a los niños, es necesario sospechar de dos causas principales de error: la sugestión y las fallas de atención. Este no es el momento de hablar del primer punto. En cuanto al segundo, las fallas de atención son tan importantes que siempre es necesario sospechar de ellas cuando se obtiene un resultado negativo. En ese momento se deben suspender los experimentos y retomarlos en una ocasión más favorable, repitiéndolos 10, 20 veces, con enorme paciencia. De hecho, es frecuente que los niños estén poco dispuestos a prestarle atención a los experimentos que no son entretenidos, y es inútil esperar que se les pueda hacer que presten mayor

atención amenazándolos con un castigo. Sin embargo, con estratagemas particulares, en ocasiones se puede dar un cierto atractivo al experimento (Binet, 1895, citado en Pollack, 1971).

Es interesante comparar las prácticas actuales de prueba —que llegan al grado de especificar el fraseo exacto que debe utilizar el examinador— con el consejo de Binet acerca de ejercitar una paciencia casi interminable y utilizar trucos entretenidos cuando se somete a prueba a los niños.

BINET Y EL EXAMEN DE LOS PROCESOS MENTALES SUPERIORES

En 1896, Binet y su ayudante en la Sorbona, Victor Henri, publicaron una revisión fundamental de los trabajos alemanes y estadounidenses sobre diferencias individuales. En este artículo, que posee gran importancia histórica, discutieron que la inteligencia podía medirse mejor a través de los procesos psicológicos superiores que con procesos sensoriales elementales, como el tiempo de reacción. Después de varios inicios en falso, Binet y Simon llegaron en un momento dado al sencillo formato de sus escalas de 1905, que se analizan más adelante.

El carácter de la escala de 1905 le debe mucho a una prueba anterior elaborada por el Dr. Blin (1902) y a su alumno M. Damaye. Estos investigadores habían intentado mejorar el diagnóstico del retraso mental a través del uso de una batería de evaluación en 20 áreas como lenguaje hablado, conocimiento de las partes del cuerpo, obediencia a órdenes simples; nombrado de objetos comunes y capacidad para leer, escribir y hacer cálculos aritméticos sencillos. Binet criticó la escala por considerarla demasiado subjetiva, por tener reactivos que reflejaban la educación formal y por utilizar un formato de sí o no en muchas preguntas (DuBois, 1970). Pero se mostró sumamente impresionado por la idea de utilizar una batería de pruebas, una característica que adoptó en sus escalas de 1905.

En 1904, el ministro de Educación Pública en París *(Minister of Public Instruction)* estableció una comisión para decidir sobre las medidas educativas que deberían emprenderse con aquellos niños que no podían beneficiarse de la educación regular. La comisión concluyó que deberían utilizarse exámenes médicos y educativos para identificar a aquellos niños que no podían aprender a través de métodos regulares. Lo que es más, se determinó que se debería retirar a estos niños de las clases regulares y darles instrucción especial, adecuada a su habilidad intelectual más limitada. Este fue el principio de las clases de educación especial.

Era evidente la necesidad de un medio para seleccionar a los niños para tal asignación especial y se le solicitó a Binet y a su colaborador Simon que desarrollaran una herramienta práctica dirigida únicamente a dicho propósito. Así surgió la primera escala formal para evaluar la inteligencia de los niños.

Goodenough (1949) ha resumido las cuatro maneras en que la escala de 1905 difería con respecto a aquéllas construidas antes.

1. No afirmaba medir precisamente ninguna facultad en particular. Más bien, se dirigía a evaluar el desarrollo mental general del niño con un grupo heterogéneo de tareas. Así, la meta no era la medición, sino la clasificación.
2. Era una prueba breve y práctica.Tomaba menos de una hora para su aplicación y requería pocos elementos en cuanto a equipo.
3. Medía de manera directa lo que Binet y Simon consideraban como el factor esencial de la inteligencia —el juicio práctico— en lugar de perder el tiempo con capacidades de menor nivel que implicaran elementos sensoriales, motores o perceptuales. Asumían una perspectiva práctica de la inteligencia.

Nos parece que en la inteligencia existe un elemento fundamental, cuya falta o alteración tiene la mayor importancia para la vida práctica; éste es el juicio, conocido también como sentido

común, sentido práctico, iniciativa o facultad para adaptarse. Juzgar bien, comprender bien y razonar bien son los manantiales de la inteligencia (Binet y Simon, 1905; según la traducción en Fancher, 1985).

4. Los reactivos estaban distribuidos según su nivel aproximado de dificultad, en vez de en función del contenido. Se realizó una estandarización preliminar con 50 niños normales cuyas edades iban de 3 a 11 años y también con varios niños subnormales y con retraso.

Las 30 pruebas en la escala de 1905 iban de las pruebas sensoriales abiertamente simples, a las abstracciones verbales bastante complejas. Así, la escala era apropiada para evaluar la gama completa de la inteligencia —desde el retraso mental grave hasta los niveles superiores de la inteligencia dotada. La escala completa se resume en el cuadro 1–1.

Excepto por las pruebas muy sencillas que se diseñaron para la clasificación de los **idiotas** de grado muy inferior (un término diagnóstico muy desafortunado que se ha abandonado desde entonces), las pruebas tenían importantes cargas hacia las habilidades verbales, lo cual refleja el alejamiento de Binet con respecto a la tradición de Galton.

Un punto interesante que con frecuencia pasan por alto los alumnos de psicología en la actualidad, es que Binet y Simon no ofrecieron en su escala de 1905 un método preciso para llegar a una puntuación total. Sería bueno recordar que su propósito era la clasificación, no la medición, y que su motivación era completamente humanitaria, es decir, la de identificar a aquellos niños que necesitaban asignarse a educación especial. Según normas contemporáneas, es difícil aceptar la confusión inherente a dicho enfoque, pero ello puede reflejar una inclinación moderna hacia la cuantificación más que una debilidad por parte de la escala de 1905. De hecho, su escala fue popular entre los educadores en París. Incluso en ausencia de una cuantificación precisa, el enfoque fue exitoso en la selección de candidatos para clases especiales.

LAS ESCALAS REVISADAS Y EL ADVENIMIENTO DEL CI

En 1908, Binet y Simon publicaron una revisión de la escala de 1905. En la escala anterior, más de la mitad de los reactivos habían sido diseñados para los individuos con retraso muy notable; sin embargo, las principales decisiones diagnósticas implicaban a niños mayores y a personas con un intelecto limítrofe. Para remediar este desequilibrio, se abandonó la mayor parte de los reactivos muy simples y se añadieron nuevos al extremo superior de la escala. La escala de 1908 tenía 58 problemas o pruebas, casi el doble del número de 1905. Se añadieron nuevas pruebas, muchas de las cuales todavía se utilizan hoy día: reconstrucción de oraciones en desorden, copiado de un rombo y realización de una secuencia de tres órdenes. Algunos de los reactivos consistían en absurdos que los niños tenían que detectar y explicar. Uno de dichos reactivos era divertido para los niños franceses: "Se halló el cuerpo de una desafortunada niña, cortado en 18 trozos. Se piensa que la niña se suicidó". Sin embargo, este reactivo era muy perturbador para algunos individuos estadounidenses, lo cual demuestra la importancia de los factores culturales en la inteligencia (Fancher, 1985).

La principal innovación de la escala de 1908 era la introducción del concepto de nivel mental. Las pruebas se habían estandarizado con cerca de 300 niños normales entre 3 y 13 años de edad. Esto permitió que Binet y Simon ordenaran las pruebas según el nivel de edad en el que por lo común se aprobaban. Cualesquiera reactivos que aprobaran de 80 a 90% de los niños de tres años, se asignaba al nivel de tres años y así sucesivamente, hasta los 13 años. Binet y Simon también diseñaron un sistema aproximado de calificación donde se determinaba primero una edad basal a partir del nivel de edad en que no se fallaba en más de una prueba. Por cada cinco pruebas aprobadas a niveles por arriba del nivel basal, se concedía un año completo de nivel mental. En la medida en que no se daba crédito por años parciales de nivel mental y los diversos niveles de edad

Cuadro 1–1. Escala Binet-Simon de 1905

1. Sigue con los ojos un objeto en movimiento.
2. Toma un pequeño objeto cuando se le toca.
3. Toma un pequeño objeto cuando se le ve.
4. Reconoce la diferencia entre un cuadro de chocolate y uno de madera.
5. Encuentra e ingiere un cuadro de chocolate envuelto en papel.
6. Obedece órdenes sencillas e imita ademanes simples.
7. Señala objetos familiares nombrados; por ejemplo: "Muéstrame la taza"
8. Señala objetos representados en láminas; por ejemplo: "Coloca tu dedo en la ventana".
9. Nombra objetos representados en dibujos; por ejemplo: "¿Qué es esto?" [el examinador señala un dibujo de un signo].
10. Compara dos líneas de longitud notablemente desigual.
11. Repite tres dígitos hablados.
12. Compara dos pesos.
13. Muestra susceptibilidad a la sugestión.
14. Define palabras comunes según su función.
15. Repite una oración de 15 palabras.
16. Explica la diferencia entre dos objetos comunes; por ejemplo: "papel y cartón"
17. De un total de 13 objetos presentados en un tablero, menciona de memoria el mayor número posible en un lapso de 30 seg. [Esta prueba se abandonó después debido a que propiciaba demasiadas posibilidades de distracción.]
18. Reproduce de memoria dos diseños mostrados durante 10 seg.
19. Repite una serie más larga de dígitos que en el punto 11 para someter a prueba la memoria inmediata.
20. Dice en qué se parecen dos objetos; por ejemplo: "mariposa y pulga".
21. Compara dos líneas de longitud un tanto desigual.
22. Compara cinco cubos para ordenarlos por peso.
23. Indica cuál de los cinco pesos anteriores ha retirado el examinador.
24. Rima palabras; por ejemplo: "¿Qué rima con 'escuela'?"
25. Una prueba de palabras incompletas basada en aquellas propuestas por Ebbinghaus.
26. Coloca tres sustantivos; por ejemplo: "París, río, fortuna" (o tres verbos) dentro de una oración.
27. Responde a 25 preguntas abstractas (comprensión); por ejemplo: "Cuando una persona te ha ofendido y viene a ofrecer disculpas, ¿qué deberías hacer?"
28. Invierte las manecillas de un reloj.
29. Después de doblar y cortar papel, dibuja la forma de los orificios resultantes.
30. Define palabras abstractas estableciendo la diferencia entre ellas; por ejemplo: "hastío y cansancio".

Fuente: Basado en las traducciones que aparecen en Jenkins y Paterson (1961) y Jensen (1980).

tenían desde 3 hasta 8 pruebas, el método dejaba mucho que desear.

En 1911 apareció una tercera revisión de las escalas Binet-Simon. Ahora, cada nivel de edad tenía exactamente cinco pruebas. La escala también se extendió hasta el rango adulto. Y, con cierta renuencia, Binet introdujo nuevos métodos de calificación que concedían una quinta parte de un año por cada subprueba aprobada por encima del nivel basal. En sus escritos, Binet enfatizó en gran medida que no debería tomarse demasiado en serio el nivel mental exacto

del niño como una medida absoluta de la inteligencia.

Sin embargo, la idea de obtener un nivel mental fue un desarrollo monumental que habría de influir el carácter de las pruebas de inteligencia a lo largo del siglo XX. Después de unos cuantos meses, lo que Binet llamó nivel mental se traducía como edad mental. Y los examinadores en todas partes, incluyendo Binet mismo, comparaban la edad mental de un niño con su edad cronológica. Así, un niño de nueve años que funcionaba al nivel mental (o edad mental) de un niño de seis años, tenía un retraso de tres años. Muy poco tiempo después, Stern (1912) señaló que un retraso de tres años tenía diferentes significados, dependiendo de la edad. Un niño de cinco años que funciona a un nivel de dos años de edad tiene una alteración mayor que uno de 13 años que funciona a un nivel de 10 años. Stern sugirió que un cociente de inteligencia calculado a partir de la división de la edad mental entre la edad cronológica daría una mejor medida del funcionamiento relativo de una persona en comparación con individuos de su misma edad.

En 1916, Terman y sus colaboradores en Stanford revisaron las escalas Binet-Simon y produjeron la Stanford-Binet, una exitosa prueba que se analizará en un capítulo posterior. Terman sugirió multiplicar el Cociente de Inteligencia por 100 para eliminar las fracciones; también fue la primera persona en utilizar la abreviatura CI. Así nació uno de los conceptos más populares y polémicos de la historia de la psicología. Binet murió en 1911, antes de que el CI se generalizara en las pruebas estadounidenses, de modo que nunca se sabrá lo que habría pensado de este nuevo desarrollo basado en sus escalas. Sin embargo, su colaborador Simon afirmó posteriormente que el concepto de CI era una "traición" a los objetivos originales de su escala (Fancher, 1985, p. 104) y se puede suponer, con base en las preocupaciones humanistas de Binet, que habría tenido una opinión similar.

RESUMEN

1. Para bien o para mal, los resultados de una prueba psicológica tienen el poder de alterar vidas. Una revisión de las tendencias históricas es crucial si se desea comprender la influencia contemporánea de las pruebas psicológicas.
2. Las formas rudimentarias de las pruebas datan del año 2200 a.C. en China. Los emperadores chinos utilizaban extenuantes exámenes por escrito para seleccionar a los funcionarios públicos.
3. Desde mediados hasta finales del siglo XIX, varios médicos y psiquiatras desarrollaron procedimientos estandarizados para revelar la naturaleza y grado de los síntomas en personas con enfermedad mental y lesión cerebral. Por ejemplo, en 1885, Hubert von Grashey desarrolló el precursor del cilindro mnemónico para someter a prueba la habilidad de reconocimiento visual de pacientes con lesión cerebral.
4. Las pruebas psicológicas modernas le deben su comienzo a la era de los instrumentos de bronce en psicología que floreció en Europa durante finales del siglo XIX. Mediante el examen de los umbrales sensoriales y tiempos de reacción, los pioneros en el desarrollo de pruebas, como Sir Francis Galton, demostraron que era posible medir la mente de manera objetiva y replicable.
5. Wilhelm Wundt fundó el primer laboratorio psicológico en Leipzig, Alemania, en 1879. Incluida entre sus primeras investigaciones se encuentra su intento, realizado en 1862, de medir la velocidad de pensamiento con un medidor del pensamiento, un péndulo calibrado con agujas que sobresalían a cada lado.
6. La primera referencia a las pruebas mentales apareció en 1890 en un artículo clásico de James McKeen Cattell, un psicólogo

estadounidense que había estudiado con Galton. Cattell importó a EUA el enfoque de los instrumentos de bronce.

7. Uno de los alumnos de Cattell, Clark Wissler, mostró que las medidas de tiempo de reacción y de discriminación sensorial no se correlacionaban con las calificaciones universitarias, con lo cual se redirigió al movimiento de las pruebas mentales en dirección opuesta a los instrumentos de bronce.
8. A finales del siglo XIX, un humanismo reciente hacia los individuos con retraso mental, reflejado en el trabajo diagnóstico y terapéutico de los médicos franceses Esquirol y Seguin, ayudó a crear la necesidad de las primeras pruebas de inteligencia.
9. Alfred Binet, quien habría de inventar la primera prueba verdadera de la inteligencia, comenzó su carrera estudiando la parálisis histérica con el neurólogo francés Charcot. La afirmación de Binet de que el magnetismo podía curar la histeria fue refutada, para su vergüenza. Poco después cambió de intereses y llevó a cabo estudios sensoperceptuales con ayuda de sus hijos.
10. En 1905, Binet y Simon elaboraron la primera prueba útil de inteligencia en París, Francia. Su sencilla medida de 30 reactivos dirigida a las funciones mentales superiores ayudó a identificar a los escolares que no se podían beneficiar de la educación regular. Curiosamente, no había un método para calificar la prueba.
11. En 1908, Binet y Simon publicaron una escala revisada con 58 reactivos, la cual incorporaba el concepto de nivel mental. En 1911 apareció una tercera revisión de las escalas Binet-Simon. Ahora, cada nivel de edad contenía exactamente cinco pruebas; la escala se extendió hasta el rango adulto.
12. En 1912, Stern propuso dividir la edad mental entre la edad cronológica para obtener un Cociente de Inteligencia. En 1916, Terman sugirió multiplicarlo por 100 para eliminar las fracciones. Así nació el concepto de CI.

TEMA 1B
Primeras pruebas en EUA

Primeros usos y abusos de las pruebas en EUA
Invención de las pruebas no verbales a principios del siglo XX
La Stanford-Binet: primer pilar del CI
Pruebas grupales y la clasificación de los reclutas del ejército en la primera Guerra Mundial
Primeras pruebas educativas
Desarrollo de las pruebas de aptitud
Pruebas de personalidad y vocacionales después de la primera Guerra Mundial
Orígenes de las pruebas proyectivas
Desarrollo de los Inventarios de Intereses
Resumen de los principales hitos en la historia de las pruebas
Resumen

Las escalas Binet-Simon ayudaron a resolver un dilema social práctico; a saber, cómo identificar a los niños que necesitaban educación especial. Con esta exitosa aplicación de una prueba mental, los psicólogos se percataron de que sus inventos podían tener importancia práctica para diferentes sectores de la sociedad. Casi de inmediato, los psicólogos en EUA adoptaron una perspectiva utilitaria. Muchos vieron en las pruebas de inteligencia una respuesta confiable y objetiva a los problemas sociales percibidos, como la identificación de los inmigrantes con retraso mental y la clasificación rápida y precisa de los reclutas del ejército.

La cuestión sobre si estas primeras pruebas resolvieron en realidad los problemas sociales —o tan sólo los exacerbaron— es un problema muy controvertido que se trata más adelante. Algo que sí es seguro es que la profusión de pruebas elaboradas a principios de este siglo ayudó a moldear el carácter de las contemporáneas. Una revisión de estas tendencias históricas ayudará a comprender la naturaleza de las pruebas modernas y a un mejor reconocimiento de los aspectos sociales que dichas pruebas han hecho surgir.

PRIMEROS USOS Y ABUSOS DE LAS PRUEBAS EN EUA

PRIMERA TRADUCCIÓN DE LA ESCALA BINET-SIMON

En 1906, Henry H. Goddard fue contratado por la *Vineland Training School*, en Nueva Jersey, para investigar la clasificación y educación de niños "débiles mentales". Pronto se percató de que se requeriría un instrumento diagnóstico y, por tanto, se sintió complacido de leer acerca de la escala Binet-Simon de 1908. Rápidamente se dio a la tarea de traducir la escala haciendo cambios de importancia menor, de modo que pudiera aplicarse a los niños estadounidenses (Goddard, 1910a).

Goddard (1910b) examinó a 378 residentes del centro Vineland y los clasificó con base en el diagnóstico y la edad mental. Clasificó a 73 residentes como **idiotas** debido a que su edad mental era de dos años o menos; a 205 se les denominó **imbéciles**, con edades mentales de 3 a 7 años; y a 100 como **débiles mentales**, con edades mentales de 8 a 12 años. Es útil señalar que los términos originalmente neutros y descriptivos para denominar los niveles de retraso mental —idiota, imbécil y débil mental— han ingresado al léxico cotidiano como palabras peyorativas. De hecho, Goddard hizo su propia contribución al acuñar el término diagnóstico **morón** (del griego *moronia*, que significa "necio").

Goddard (1911) también sometió a 1 547 niños normales a las pruebas traducidas de las escalas Binet-Simon. Consideró que los niños cuya edad mental se encontraba cuatro o más años por debajo de su edad cronológica eran débiles mentales —éstos constituyeron 3% de su muestra. Si se estima que todos estos niños se encontraban fuera de las instituciones para individuos con retraso, 3% es una tasa alarmante de "deficiencia mental". Goddard (1911) tenía la opinión de que se debería segregar a estos niños para evitar que "contaminaran a la sociedad". Estos primeros estudios azuzaron la curiosidad de Goddard acerca de los ciudadanos "débiles mentales" y sobre las cargas sociales que éstos imponían. También se ganó la reputación de uno de los principales expertos en el uso de pruebas de inteligencia para identificar a personas con discapacidades intelectuales. Sus talentos pronto tuvieron una fuerte demanda.

LA BINET-SIMON Y LA INMIGRACIÓN

En 1910, el comisionado de inmigración invitó a Goddard a la isla Ellis para que le ayudara a realizar un examen más preciso de los inmigrantes. A principios del siglo XX habían empezado a correr una serie de mitos oscuros y premonitorios acerca del retraso mental y la inmigración:

> Se creía que los débiles mentales eran seres degenerados responsables de muchos, si no es que de todos, los problemas sociales; se reproducían a una tasa alarmante y amenazaban la aptitud biológica general de la nación, y su número se incrementaba debido a la llegada de "nuevos" inmigrantes indeseables provenientes de países del sur y este de Europa, que habían estado suplantando a los "viejos" inmigrantes de las áreas norte y oeste de Europa (Gelb, 1986).

Al principio, a Goddard no le preocupaba la supuesta amenaza de la debilidad mental que representaban los inmigrantes. Escribió sobre la inexistencia de datos estadísticos adecuados y que las opiniones generales acerca de los grandes porcentajes de inmigrantes con defectos mentales eran "una sobrestimación exagerada" (Goddard, 1912). Sin embargo, a partir de visitas repetidas a la isla Ellis llegó a convencerse de que las tasas de debilidad mental eran mucho más altas que las estimadas por los médicos del personal del servicio de inmigración. Dentro de ese mismo año revirtió por completo sus opiniones y solicitó fondos al Congreso para contratar a expertos entrenados en el uso de pruebas de inteligencia para formar parte del personal de la isla Ellis. En el siguiente decenio, Goddard se volvió un apóstol del uso de las pruebas de inteligencia para identificar a los inmigrantes con debilidad mental. Aunque escribió que las tasas de inmigrantes con ese

problema eran "alarmantes", no se adhirió a la petición popular de restringir la inmigración (Gelb, 1986).

Con frecuencia, la historia de Goddard y su preocupación por "la amenaza de la debilidad mental", como lo ha expresado satíricamente Gould (1981), se ha ignorado o se le ha restado importancia en los libros sobre pruebas psicológicas. La mayor parte de los textos sobre pruebas no se refieren o mencionan a Goddard en absoluto. Los pocos libros que hablan de él expresan en general que Goddard: ". . . utilizó las pruebas en instituciones para individuos con retraso. . .", lo cual es con toda seguridad una subestimación. En su influyente libro *History of Psychological Testing*, DuBois (1970) tiene una fotografía de Goddard, pero le dedica menos de una línea dentro del texto.

El hecho es que Goddard fue uno de los psicólogos estadounidenses con mayor influencia a principios del siglo XX. Por tanto, cualquier persona cuidadosa debe preguntarse por qué tantos autores contemporáneos han ignorado o despreciado a quien fue el primero en traducir y aplicar las pruebas de Binet en EUA. Aquí se intentará responder esta pregunta, con base, en parte, en los escritos originales de Goddard, pero también en función de la crítica de Gould (1981) acerca de los voluminosos textos de Goddard sobre retraso mental y pruebas de inteligencia. También se hace referencia al retrato más compasivo de Goddard que nos proporciona Gelb (1986).

Quizá Goddard ha sido ignorado en los libros de texto debido a que era un estricto creyente en la herencia, la cual concebía a la inteligencia en términos mendelianos simples. Sin duda, su petición de establecer una colonia de "morones", de modo que se restringiera su reproducción, le ha ganado también la desaprobación contemporánea. Y su insistencia en que conductas indeseables como criminalidad, alcoholismo y prostitución eran el producto del retraso mental hereditario, tampoco se ajusta con la postura ambientalista moderna.

No obstante, la razón más probable por la que los autores modernos han ignorado a Goddard, es que ejemplificaba a gran número de los primeros psicólogos más destacados que participaron en el obvio uso incorrecto de las pruebas de inteligencia. En sus esfuerzos por demostrar que cada día ingresaban a EUA gran número de inmigrantes con retraso, Goddard envió a sus ayudantes a la isla Ellis para aplicar su traducción inglesa de las pruebas Binet-Simon a inmigrantes recién llegados. Las pruebas se aplicaban a través de un traductor, poco después de que las personas habían arribado. Se puede suponer que muchos de ellos se encontraban atemorizados, confundidos y desorientados; por tanto, una prueba diseñada en francés y traducida luego al inglés, se retraducía posteriormente al yiddish, húngaro, italiano o ruso; se le aplicaba a granjeros y obreros confundidos que hacía poco habían soportado una travesía por el Atlántico y se le interpretaba según las normas originales francesas.

¿Qué encontró Goddard y cuáles fueron sus conclusiones acerca de los resultados? En pequeñas muestras de inmigrantes (22 a 50), sus ayudantes encontraron que 83% de los judíos, 80% de los húngaros, 79% de los italianos y 87% de los rusos eran débiles mentales; es decir, se encontraban por debajo de los 12 años de edad en las escalas Binet-Simon (Goddard, 1917). Su interpretación de estos datos es, según el caso, de un escepticismo cauto y después de un alarmismo provocativo. En una parte del informe afirma que su estudio "no hace ninguna determinación acerca del porcentaje real, incluso dentro de estos grupos, en cuanto a débiles mentales". Sin embargo, después expresa que sus cifras sólo necesitan revisión en cuanto a "una cantidad relativamente pequeña", con el fin de encontrar los porcentajes reales de débiles mentales entre los grupos de inmigrantes. Lo que es más, concluye que la inteligencia del inmigrante promedio es baja, "quizá a nivel de morón", pero después cita la privación ambiental como el principal culpable. De modo simultáneo, Goddard parece favorecer la deportación para los inmigrantes con un CI bajo, pero también expresa la perspectiva humanitaria de que se podría utilizar a "trabajadores morones" si sólo "somos suficientemente sabios como para entrenarlos de manera apropiada".

Existe mucho más con respecto a la era de Goddard de las primeras pruebas de inteligencia y se alienta al lector interesado a consultar a Gould (1981) y Gelb (1986). El punto más importante que se quiere destacar en este caso es que —como muchos de los primeros psicólogos— las perspectivas académicas de Goddard estuvieron bajo la influencia de las ideologías sociales de su tiempo. Por último, Goddard fue un erudito complejo que refinó y contradijo sus opiniones profesionales en numerosas ocasiones. Un ejemplo irónico de ello es que, después de que se había causado el daño y sus escritos habían ayudado a restringir la inmigración, se retractó (1928), concluyendo que la debilidad mental no era incurable y que no era necesario recluir a los débiles mentales en instituciones.

El capítulo sobre Goddard en la historia de las pruebas sirve para recordar que incluso las personas bien intencionadas, que operan dentro de normas sociales aceptadas en términos generales, pueden hacer uso incorrecto de las pruebas psicológicas. Siempre se debe tener presente que la "ciencia" desinteresada puede vincularse con las metas de una ideología social perniciosa.

INVENCIÓN DE LAS PRUEBAS NO VERBALES A PRINCIPIOS DEL SIGLO XX

Debido al énfasis de las escalas Binet-Simon en las habilidades verbales, muchos psicólogos se dieron cuenta de que este nuevo instrumento de medición no era totalmente apropiado para personas no angloparlantes, analfabetas o con problemas de audición y lenguaje. Por tal motivo, después de la traducción de Goddard para la Binet-Simon de 1908, surgió un torrente de escalas de ejecución. Aquí sólo se proporcionará una breve cronología de las pruebas no verbales. El lector interesado debe consultar a DuBois (1970). En su listado de las primeras pruebas de ejecución, el lector encontrará con toda seguridad muchos instrumentos y subpruebas que aún están en uso.

La primera de las medidas de ejecución fue el tablero Seguin de figuras, una base vertical con depresiones en las que se pueden colocar 10 bloques de diversas formas. Seguin lo utilizó como medio de entrenamiento para individuos con retraso, pero después Goddard lo empleó como prueba y posteriormente R. H. Sylvester (1913) lo estandarizó. En la batería de pruebas neuropsicológicas Halstead-Reitan aún se utiliza un tablero idéntico, para lo cual se vendan los ojos del sujeto (Reitan y Wolfson, 1985).

Knox (1914) diseñó varias pruebas de ejecución para utilizarlas con los inmigrantes en la isla Ellis. Sus pruebas no requerían ningún tipo de respuesta verbal de parte de los sujetos. El examinador demostraba de manera no verbal las tareas para garantizar que los examinandos comprendían las instrucciones. Dentro de esta prueba se incluían un rompecabezas simple de madera (al que Knox denominaba prueba para "morones") y la misma prueba de sustitución de símbolos y dígitos que se encuentra ahora en la mayor parte de las escalas Wechsler de inteligencia.

Varias de las primeras pruebas de ejecución son dignas de una mención breve debido a que han sobrevivido hasta la actualidad ya corregidas. Pintner y Paterson (1917) inventaron una escala de pruebas de ejecución en 15 partes que utilizaban varios tableros de figuras, rompecabezas y pruebas de ensamble de objetos. Esta última —que consiste en reconstruir partes de cartón recortadas de un objeto común, como un caballo— es un pilar de varias pruebas contemporáneas de inteligencia. La de Diseños con Cubos de Kohs (Kohs, 1920), que requería que la persona ensamblara cubos pintados para reproducir un patrón, es muy conocida para cualquier examinador moderno que utilice las escalas Wechsler. La Prueba Porteus de Laberintos (Porteus, 1915) consiste en una serie graduada de laberintos en los cuales el sujeto debe evitar los callejones sin salida, mientras traza un camino desde el principio hasta el fin. Éste es un buen instrumento que todavía está disponible en la actualidad, pero se le utiliza muy poco.

LA STANFORD-BINET: PRIMER APOYO DEL CI

Aunque fue Goddard quien tradujo por primera vez las escalas Binet en EUA, fue un profesor de Stanford, Lewis M. Terman (1857-1956), quien popularizó las pruebas de CI revisadas de las escalas Binet en 1916. La nueva Stanford-Binet (así se le llamó) no sólo fue una extensión de las primeras escalas Binet, sino una revisión sustancial. Entre los muchos cambios que condujeron al prestigio incuestionable de la Stanford-Binet, se encontraba el uso del ahora familiar CI para expresar los resultados de la prueba. Se aumentó a 90 el número de los reactivos y la nueva escala era adecuada para individuos con retraso mental, niños y adultos tanto normales como "superiores". Además, la Stanford-Binet tenía instrucciones claras y bien organizadas para aplicación y calificación. Se tuvo gran cuidado en garantizar una muestra representativa de sujetos para la estandarización de la prueba. Como señala Goodenough (1949): "La publicación de la Revisión Stanford señaló el fin del periodo inicial de experimentación e incertidumbre. De una vez por todas, el examen de la inteligencia se había colocado sobre una base firme."

La Stanford-Binet constituyó por decenios la norma para las pruebas de inteligencia. Las nuevas pruebas se validaban siempre en términos de sus correlaciones con esta medida, que continuó a lo largo de sus revisiones en 1937 y 1960, momento en el cual las escalas Wechsler (Wechsler, 1949, 1955) habían empezado a competir con ella. La última revisión de la Stanford-Binet se concluyó en 1986. Esta prueba y las escalas Wechsler se analizan con detalle en un capítulo posterior. Es necesario mencionar aquí que las escalas Wechsler se volvieron una alternativa muy popular para la Stanford-Binet, principalmente debido a que proporcionaban más de una puntuación de CI. Además del CI Total, las escalas Wechsler proporcionaban de 10 a 12 puntuaciones de subprueba y un CI Verbal y uno de Ejecución. En contraste, las versiones anteriores de la Stanford-Binet sólo proporcionaban una puntuación general de resumen, el CI global.

PRUEBAS GRUPALES Y CLASIFICACIÓN DE RECLUTAS DEL EJÉRCITO EN LA PRIMERA GUERRA MUNDIAL

Dada la tendencia estadounidense por la eficiencia, era natural que los investigadores buscaran pruebas mentales de grupo que complementaran a las pruebas de inteligencia individuales importadas de Francia, las cuales consumían relativamente demasiado tiempo. Entre los primeros en desarrollar pruebas grupales se encontraba Pyle (1913), quien publicó normas para niños en edad escolar para una batería que incluía medidas tan trilladas como capacidad de memoria, sustitución de símbolos y dígitos y asociación de palabras orales (escribir rápidamente palabras en respuesta a una palabra estímulo). Pintner (1917) revisó y aumentó la batería de Pyle, añadiéndole una prueba de eliminación con límite de tiempo, en la que el niño tachaba la letra a cada vez que aparecía en un texto.

Pero las pruebas grupales tardaron en alcanzar la popularidad, en parte debido a que las primeras versiones todavía tenían que calificarse de manera laboriosa a mano. La idea de una prueba completamente objetiva con una clave de calificación simple era inconsistente con pruebas como memoria lógica, para cuya calificación se requería el juicio del examinador. Lo más sorprendente de todo —al menos para cualquiera que haya pasado algún tiempo con un alumno en una escuela en EUA— es que las preguntas de opción múltiple todavía no se utilizaban de manera general.

El paso lento de los avances en cuanto a pruebas grupales se elevó de modo notable cuando EUA entró a la primera Guerra Mundial en 1917. Fue entonces que Robert M. Yerkes, un conocido profesor de psicología en Harvard, convenció al gobierno y al ejército de que sus 1.75 millones de reclutas deberían pasar por una prueba de inteligencia con propósitos de clasificación y asignación (Yerkes, 1919). De inmediato, al ser comisionado dentro del ejército con el grado de coronel, Yerkes

organizó un Comité de Examen de Reclutas, que se reunió en la escuela Vineland, en Nueva Jersey, para desarrollar el nuevo grupo de pruebas para la evaluación de los reclutas. Yerkes presidió dicho comité y otros miembros famosos incluyeron a Goddard y Terman.

De esta colaboración surgieron dos pruebas grupales: la Army Alfa y la Army Beta. Sería difícil sobrestimar la influencia de dichas pruebas sobre las subsecuentes de inteligencia. Su formato y contenido inspiraron los desarrollos de pruebas grupales e individuales durante los siguientes decenios. Se analizará con cierto detalle éstas, de modo que el lector pueda apreciar su influencia sobre las pruebas modernas de inteligencia.

EXÁMENES ARMY ALFA Y BETA

La Army Alfa se basó en el trabajo aún no publicado de Otis (1918) e incluía ocho pruebas con carga verbal para reclutas que tenían desde un funcionamiento promedio hasta uno superior. Las ocho pruebas eran: 1) seguimiento de instrucciones orales; 2) razonamiento aritmético; 3) juicio práctico; 4) pares de sinónimos-antónimos; 5) oraciones desordenadas; 6) llenado de series de números; 7) analogías y 8) información. La figura 1–1 muestra algunos reactivos típicos del examen Army Alfa.

La Army Beta era una prueba grupal no verbal diseñada para individuos analfabetas y reclutas cuyo primer idioma no era el inglés.

SEGUIMIENTO DE INSTRUCCIONES ORALES

Marque una cruz en el primer y tercer círculos:

O O O O O

RAZONAMIENTO ARITMÉTICO

Resuelva cada problema:

¿Cuántos son 5 hombres más 10 hombres? Respuesta ()

Si 3½ toneladas de carbón cuestan $21, ¿cuánto costarán 5½ toneladas? Respuesta ()

JUICIO PRÁCTICO

¿Por qué las montañas altas están cubiertas de nieve? Porque

- ☐ están cerca de las nubes
- ☐ el sol brilla en raras ocasiones sobre ellas
- ☐ el aire es frío en ese lugar

PARES DE SINÓNIMOS–ANTÓNIMOS

¿Las siguientes palabras son iguales u opuestas?

generosidad–donación ¿iguales? u ¿opuestas?

acumular–disipar ¿iguales? u ¿opuestas?

ORACIONES DESORDENADAS

¿Las siguientes palabras pueden reordenarse para formar una oración?

envidia malos maldad rasgos son y ¿cierto? o ¿falso?

COMPLETAR DE SERIES DE NÚMEROS

Complete la serie: 3 6 8 16 18 36

ANALOGÍAS

¿Cuál opción completa la analogía?

lágrimas–pena :: risa– gozo sonrisa niñas mueca

granero–trigo :: biblioteca– escritorio libros papel bibliotecaria

INFORMACIÓN

Elija la mejor alternativa:

El páncreas se encuentra en el abdomen cabeza hombro cuello

La Batalla de Gettysburg se peleó en 1863 1813 1778 1812

Nota: Los examinados recibían instrucciones orales para cada subprueba.

Figura 1–1. Reactivos muestra del examen Army Alfa.
Fuente: Reproducido de Yerkes, R. M. (ed.) (1921). *Psychological examining in the United States Army. Memoirs of the National Academy of Sciences*, Volumen 15. Con autorización de la National Academy of Sciences, Washington, DC.

Incluía varias pruebas visoperceptuales y motoras, como seguir un camino en un laberinto y visualizar el número correcto de cubos representados en un dibujo tridimensional. La figura 1–2 presenta las demostraciones de pizarrón para las ocho partes del examen Army Beta.

A fin de dar cabida a los sujetos analfabetas y a los inmigrantes recientes que no comprendían el inglés, Yerkes instruyó a los examinadores para que utilizaran métodos que eran en gran medida pictóricos y expresivos, para explicarles las pruebas a los futuros reclutas

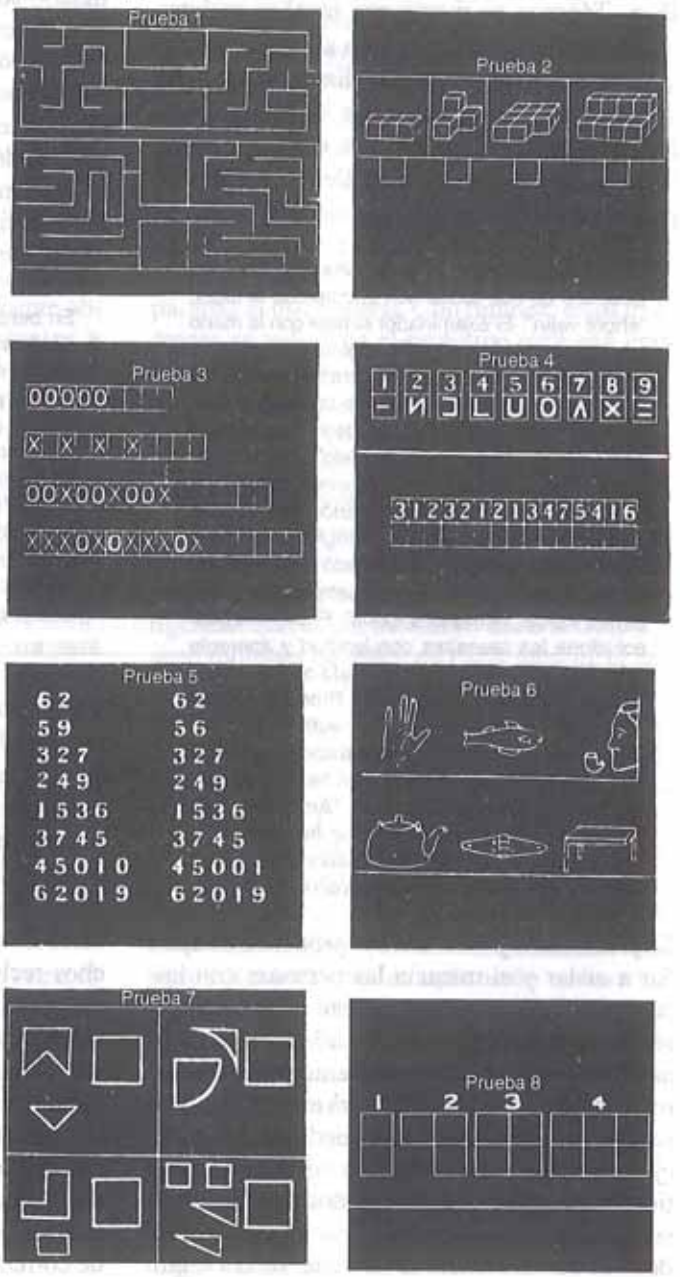


Figura 1–2. Demostraciones de pizarrón de las ocho partes del examen Beta **Fuente:** Reproducido de Yerkes, R.M. (ed.) (1921). *Psychological examining in the United States Army. Memoirs of the National Academy of Sciences,* Volumen 15. Con autorización de la National Academy of Sciences, Washington, DC.

del ejército. El examinador y su ayudante se encontraban de pie sobre una plataforma al frente de la clase y llevaban a cabo pantomimas para explicar cada una de las ocho pruebas. Se reproducirán aquí las instrucciones exactas para una prueba, de modo que el lector pueda evaluar los probables efectos de los procedimientos de prueba sobre los resultados del examen Beta. Téngase en mente que muchos reclutas no podían ver o escuchar bien al examinador y que algunos nunca habían realizado una prueba antes. Aquí se presenta cómo los examinadores introducían la prueba seis, dibujos incompletos, a cada nuevo salón lleno de reclutas potenciales:

> "Ésta es la prueba 6. Vean. Muchos dibujos". Después de que todos han encontrado el lugar, "ahora vean". El examinador señala con la mano y le dice al demostrador: "Arréglenlo". El demostrador no hace nada, pero parece confuso. El examinador señala al dibujo de la mano y después al lugar donde falta el dedo y le dice al demostrador: "Arréglelo; arréglelo". Entonces el demostrador dibuja el dedo. El examinador dice: "Correcto". Entonces el examinador señala al pez y al lugar para el ojo y dice: "Arréglelo". Después de que el demostrador ha dibujado el ojo faltante, el examinador señala a los cuatro dibujos restantes y dice: "Arréglelos todos". El demostrador soluciona las muestras con lentitud y aparente dificultad. Cuando se ha terminado con las muestras, el examinador dice: "Bien. Prosigan. ¡Apresúrense!" Durante el curso de esta prueba los ordenanzas caminan por la habitación y localizan a los individuos que no están haciendo nada, señalan a sus páginas y dicen: "Arréglelo. Arregle todos", intentando que todos trabajen. Al final de tres minutos, el examinador dice: "¡Alto! Pero no den vuelta a la hoja" (Yerkes, 1921).

La prueba del ejército tenía el propósito de ayudar a aislar y eliminar a las personas con incompetencia mental, clasificar a los varones según su capacidad mental y asistir en la asignación de los varones competentes en los puestos de responsabilidad (Yerkes, 1921). Sin embargo, en realidad no ha quedado claro si el ejército utilizó la totalidad de las grandes cantidades de datos proporcionados por Yerkes y sus entusiastas ayudantes. Una lectura cuidadosa de sus memorias revela que Yerkes logró poco más que producir los testimoniales favorables de los oficiales de alto rango. En lo fundamental, sus memorias dicen que el ejército pudo haber ahorrado millones de dólares e incrementado su eficacia si se hubieran empleado los datos de las pruebas.

En cierto sentido, las montañas de datos de prueba tuvieron ese impacto práctico tan limitado sobre la eficacia del ejército, debido a la resistencia de la mentalidad militar ante la innovación científica. Sin embargo, también es cierto que los oficiales de mayor rango en el ejército tenían buenas razones para dudar de la validez de los resultados de prueba. Por ejemplo, un memorándum interno describió el uso de la pantomima en las instrucciones para el examen no verbal Beta como:

> En beneficio de poder comparar los resultados de diversos campamentos, los examinadores recibieron la orden de seguir una cierta serie detallada y específica de extravagancias de ballet, que no sólo tuvieron el mérito de ser totalmente incomprensibles y no relacionadas con la prueba mental, sino que también se prestaron a una atmósfera mística sumamente confusa y distractora para todo el desempeño, con lo que, en efecto, se impidió todo acercamiento a la actitud en la que debe estar el sujeto mientras se examina su alma (citado en Samelson, 1977).

Además, las condiciones de prueba dejaron mucho que desear, con oleadas de reclutas que se conducían por una puerta, se sometían a prueba y virtualmente se sacaban a empujones por el otro lado. Decenas de miles de reclutas recibieron un cero literal en muchas subpruebas, no a causa de que fueran retrasados, sino porque no podían desentrañar las instrucciones para estos instrumentos nuevos y enigmáticos. Muchos reclutas se quedaban dormidos mientras los examinadores daban estas instrucciones esotéricas y misteriosas a través de pantomima.

En cuanto al lado positivo, las pruebas del ejército proporcionaron a los psicólogos una importante cantidad de experiencia en lo que se refiere a la psicometría de la construcción de pruebas. Se calcularon miles de coeficientes de correlación, incluyendo el notable uso de correlaciones múltiples, en el análisis de los

datos de prueba. La construcción de pruebas avanzó de ser un arte a convertirse en una ciencia en el curso de unos cuantos años.

LAS PRUEBAS DEL EJÉRCITO Y LAS DIFERENCIAS ÉTNICAS

Por desgracia, en ocasiones los resultados de las pruebas del ejército se utilizaron para fundamentar prejuicios sobre diversos grupos raciales y étnicos, en lugar de investigar de manera desapasionada las causas de las diferencias entre grupos. Por ejemplo, en su influyente libro *A Study of American Intelligence*, Brigham (1923) emprendió un análisis masivo de las puntuaciones Alfa y Beta de los inmigrantes nórdicos, mediterráneos y alpinos. El texto está lleno de cuadros y gráficas ostensiblemente objetivos que comparan a grupos raciales y étnicos, como una curiosa figura presentada en su libro que representa la proporción de cada muestra de inmigrantes en o por debajo del promedio de los reclutas afroestadounidenses. Brigham concluyó que este grupo, los inmigrantes mediterráneos y los inmigrantes alpinos, era intelectualmente inferior. Alertó en cuanto a que la mezcla entre razas produciría de manera invariable un deterioro de la inteligencia estadounidense. Por ejemplo, el pie de figura para una de las gráficas dice, en parte:

> Distribuciones de las puntuaciones de inteligencia de todo el grupo nórdico, combinación de los grupos mediterráneo y alpino y para los reclutas negros. El proceso de mezcla interracial no puede resultar en otra cosa que no sea un promedio de estos elementos, con el resultante deterioro de la inteligencia estadounidense (Brigham, 1923).

Siete años después, Brigham (1930) se retractó de manera franca de sus anteriores perspectivas. Citó las diferencias culturales y del lenguaje como la probable causa de las disparidades étnicas y raciales en las pruebas del ejército. Aseveró que no podían realizarse estudios comparativos de los grupos nacionales y raciales con las pruebas existentes y concluyó que sus anteriores conclusiones "no tenían fundamento" (Brigham, 1930).

PRIMERAS PRUEBAS EDUCATIVAS

Para bien o para mal, el esquema grandioso de Yerkes para someter a prueba a los reclutas del ejército ayudó a iniciar la era de las pruebas grupales. Después de la primera Guerra Mundial se acumularon las preguntas de la industria, las escuelas públicas y las universidades acerca de las potenciales aplicaciones de estas sencillas pruebas que casi cualquier persona podía aplicar y calificar (Yerkes, 1921). Los psicólogos que habían trabajado con Yerkes dejaron pronto el servicio militar y se llevaron con ellos a los campos industrial y educativo su recién encontrado concepto de las pruebas de inteligencia de lápiz y papel.

Las pruebas Army Alfa y Beta también pasaron al uso general. Con rapidez, estos exámenes se volvieron el prototipo para una gran familia de pruebas grupales e influyeron sobre el carácter de las pruebas de inteligencia, exámenes de admisión a las universidades, pruebas de rendimiento escolar y pruebas de aptitud. Para citar sólo una consecuencia específica de las pruebas del ejército, el *National Research Council*, una organización gubernamental de científicos, diseñó la Prueba Nacional de Inteligencia, que en un momento dado se aplicó a siete millones de niños en EUA durante el decenio de 1920. Así, pruebas tan conocidas como las escalas Wechsler, la Prueba de Aptitud Escolar y el Examen de Selección para Posgrado tienen en realidad raíces que llegan hasta Yerkes, Otis y las pruebas masivas de reclutas del ejército durante la primera Guerra Mundial.

El comité de Examen de Ingreso a la Universidad (CEEB, del inglés *College Entrance Examination Board*) se estableció a inicios del siglo XX para evitar la duplicación en los exámenes a los solicitantes de universidades en EUA. Los primeros exámenes habían tenido el formato de respuesta de ensayo corto, pero esto se cambió rápidamente cuando C. C. Brigham, un discípulo de Yerkes, se convirtió en secretario de la CEEB después de la primera Guerra Mundial. En 1925, el Comité Universitario decidió construir una prueba de aptitud escolar que se utilizaría en las admisiones a la universi-

dad (Goslin, 1963). Las nuevas pruebas reflejaron el formato objetivo ahora familiar de ordenamiento de oraciones, completamiento de analogías y llenado del número siguiente en una secuencia. En el decenio de 1930 se introdujo la calificación a través de máquinas, con lo cual se hizo que las pruebas grupales objetivas fueran incluso más eficaces que antes. Estas pruebas han evolucionado desde entonces hasta llegar a las actuales del Comité Universitario, en particular la Prueba de Aptitud Escolar, que se conoce ahora como Prueba de Evaluación Escolar (Scholastic Assessment Tests).

Las funciones de la CEEB se supeditaron luego al Servicio de Exámenes Educativos (ETS, del inglés *Educational Testing Service*), una organización sin fines de lucro. El ETS dirigió el desarrollo, estandarización y validación de pruebas tan conocidas como el Examen de Selección para Posgrado, la Prueba de Admisiones para la Escuela de Leyes y las Pruebas de Ingreso al Cuerpo de Paz.

Mientras tanto, Terman y sus colaboradores en Stanford se encontraban ocupados desarrollando pruebas estandarizadas de rendimiento. La *Stanford Achievement Test* (Prueba Stanford de Aprovechamiento, SAchT) fue la primera publicada en 1923; una versión moderna se encuentra actualmente en uso. Desde su inicio, la SAchT incorporó principios psicométricos modernos como la obtención de normas para subpruebas, de modo que pudiera evaluarse la variabilidad intrasujeto y la selección de una muestra muy grande y representativa de estandarización.

DESARROLLO DE LAS PRUEBAS DE APTITUD

Las pruebas de aptitud miden capacidades más específicas y delimitadas que las de inteligencia. Por tradición, estas últimas evalúan un constructo más global como la inteligencia general, aunque existen excepciones a esta tendencia que se analizarán después. En contraste, una prueba de una sola aptitud medirá sólo un dominio de capacidad, y una batería de prueba de aptitud múltiple proporcionará las puntuaciones de varias áreas de capacidad.

El desarrollo de las pruebas de aptitud se rezagó con respecto al de las pruebas de inteligencia debido a dos razones, una estadística y la otra social. El problema estadístico residía en que, con frecuencia, se necesitaba una nueva técnica, el análisis factorial, para discernir cuáles aptitudes eran primordiales y, por tanto, distinguirlas entre sí. Spearman (1904) había comenzado de manera bastante temprana con la investigación sobre este asunto, pero el método no se perfeccionó sino hasta el decenio de 1930 (Spearman, 1927; Kelley, 1928; Thurstone, 1938). Esta nueva familia de técnicas, el análisis factorial, le permitió a Thurstone concluir que había factores específicos de capacidad mental primaria como comprensión verbal, fluidez de vocabulario, facilidad numérica, capacidad espacial, memoria asociativa, velocidad perceptual y razonamiento general (Thurstone, 1938; Thurstone y Thurstone, 1941). En capítulos posteriores se dirá más sobre el tema de las pruebas de inteligencia y capacidad. El punto importante en este momento es que Thurstone y sus seguidores consideraron que, por un decir, las medidas globales de inteligencia no "desmenuzaban a la naturaleza". Como resultado, se pensó que las medidas como la Stanford-Binet no eran útiles como baterías de prueba de aptitudes múltiples para determinar las fortalezas y debilidades intelectuales de una persona.

La segunda razón para el lento crecimiento de las baterías de aptitud fue la ausencia de una aplicación práctica para dichos instrumentos refinados. No fue sino hasta la segunda Guerra Mundial que surgió una necesidad urgente de seleccionar candidatos altamente calificados para tareas muy difíciles y especializadas. Los requisitos de trabajo de los pilotos, ingenieros de vuelo y copilotos eran muy específicos y demandantes. Un estimado general de la capacidad intelectual, como el que proporcionaban las pruebas grupales de inteligencia utilizadas en la primera Guerra Mundial, no era suficiente para elegir buenos candidatos para la escuela de vuelo. Las fuerzas armadas resol-

vieron este problema desarrollando una batería especializada de aptitud con 20 pruebas que se aplicaba a los varones que aprobaban las pruebas de selección preliminares. Estas medidas resultaron invaluables para seleccionar pilotos, copilotos y bombarderos, como se reflejó en las tasas mucho más bajas de fracaso de los varones seleccionados con la batería de pruebas, en vez de con los viejos métodos (Goslin, 1963). Dichas pruebas todavía se utilizan ampliamente en las fuerzas armadas.

PRUEBAS DE PERSONALIDAD Y VOCACIONALES DESPUÉS DE LA PRIMERA GUERRA MUNDIAL

Aunque Galton, Kraepelin y otros habían utilizado métodos de evaluación tan rudimentarios como la técnica de asociación libre desde antes de principios de siglo, no fue sino hasta la primera Guerra Mundial que las pruebas de personalidad surgieron con una forma parecida a su apariencia actual. Como ha sucedido con tanta frecuencia en la historia de las pruebas, fue de nuevo una necesidad práctica lo que sirvió como ímpetu para este nuevo desarrollo. Las pruebas modernas de personalidad comenzaron cuando Woodworth intentó desarrollar un instrumento para detectar a los reclutas del ejército que fueran susceptibles a la psico-neurosis. Casi todos los inventarios y cuestionarios modernos de personalidad tienen una deuda con la *Woodworth's Personal Data Sheet* (Hoja de Datos Biográficos de Woodworth, 1919).

Este instrumento incluía 116 preguntas que el sujeto debía responder subrayando **Sí** o **No**. Las preguntas eran exclusivamente de la variedad "evidente a primera vista" y en su mayoría implicaban sintomatología bastante grave. Reactivos representativos incluían:

- ¿Las ideas pasan por tu mente de manera que no te permiten dormir?
- ¿Se te consideraba como un niño malo?
- ¿Te sientes perturbado por un sentimiento de que las cosas no son reales?
- ¿Tienes un fuerte deseo de suicidarte?

Con toda seguridad, los lectores familiarizados con el Inventario Multifásico de la Personalidad (MMPI) deben reconocer la deuda que tiene este inventario más reciente con el instrumento de Woodworth.

De su explicación acerca de cómo se desarrolló la Hoja de Datos Biográficos (Woodworth, 1951), queda claro que Woodworth tuvo gran cuidado en la selección de los reactivos. A pesar de ello, en otros sentidos este instrumento engloba una enorme dosis de credulidad psicométrica. El problema más grave reside simplemente en que un sujeto trastornado que tenga la motivación de quedar bien podría lograrlo sin que se le detectara; de la misma manera, un sujeto normal con una mentalidad de **dar una mala impresión**, podría categorizarse como incapaz para el servicio militar. Los instrumentos modernos como el MMPI han incorporado diversas escalas de validez para detectar estas tendencias de respuesta. En contraste, la Hoja de Datos Biográficos se fundamentaba en la suposición de que los sujetos serían sinceros al responder a las preguntas.

El siguiente desarrollo importante fue un inventario de neurosis, el Inventario de Personalidad de Thurstone (Thurstone y Thurstone, 1930). Después de seleccionar primero miles de reactivos que se pudieran responder con Sí-No, del inventario de Woodworth y de otras fuentes, Thurstone adaptó de manera racional los reactivos en términos de la forma en que un neurótico respondería típicamente a ellos. Como un reflejo de la inclinación de Thurstone hacia el refinamiento estadístico, este inventario fue uno de los primeros en utilizar el método de consistencia interna, donde cada posible reactivo se correlacionó con la puntuación total en una escala tentativamente identificada, para determinar si dicho reactivo pertenecía a ésta.

De la prueba Thurstone surgió *Bernreuter Personality Inventory* (Inventario de Personalidad de Bernreuter; Bernreuter, 1931). Fue un poco más refinado que su predecesor en la

medición de cuatro dimensiones de la personalidad: tendencia neurótica, autosuficiencia, introversión-extraversión y dominancia-sumisión. Una de las principales innovaciones en cuanto a construcción de pruebas fue que un solo reactivo de prueba podía contribuir a más de una escala.

El Estudio de Valores de Allport-Vernon también se publicó en 1931 (Allport y Vernon, 1931). Esta prueba fue bastante diferente de las otras en cuanto a que medía valores en lugar de psicopatología. Lo que es más, adoptó un nuevo método de calificación, el enfoque ipsativo, en el que sólo se establece una comparación de la persona consigo misma con respecto al equilibrio de importancia de seis valores básicos dados: teóricos, económicos, estéticos, sociales, políticos y religiosos. La prueba se diseñó de tal manera que se pedía a los sujetos que eligieran entre los seis valores en situaciones específicas. Como consecuencia, el promedio de las seis escalas siempre era el mismo para cada sujeto. Una debilidad en un valor se compensaba con una fortaleza en algún otro. Así, sólo tenían interés las cimas y valles relativos.

Cualquier cronología de los inventarios de autoinforme seguramente debe incluir al Inventario Multifásico de la Personalidad (MMPI; Hathaway y McKinley, 1940). Esta prueba y su revisión, el MMPI-2, se analizarán con mayor detalle después. Por ahora bastará con señalar que las escalas del MMPI se construyeron con el método del cual Woodworth fue el pionero, la contrastación de respuestas de sujetos normales y con trastornos psiquiátricos. Además, el MMPI introdujo el uso de las escalas de validez para determinar patrones de respuesta falsamente malos, falsamente buenos y aleatorios.

ORÍGENES DE LAS PRUEBAS PROYECTIVAS

El enfoque proyectivo se originó con el método de asociación de palabras, cuyo pionero fue Francis Galton a finales del siglo XIX. Galton se daba a sí mismo cuatro segundos para dar tantas asociaciones como fuera posible para una palabra estímulo y después las categorizaba como repetitivas, mediadas por imágenes o representaciones histriónicas. Esta última categoría lo convenció de que entraban en juego operaciones mentales que "están totalmente sumergidas de bajo del nivel de la conciencia". Incluso algunos historiadores han especulado que la aplicación que hizo Freud de la asociación libre como herramienta terapéutica en el psicoanálisis se derivó del artículo que publicó Galton en 1879 en la revista *Brian* (Forrest, 1974).

Wundt y Kraepelin continuaron en Alemania con el trabajo de Galton y, finalmente, Jung (1910) lo concretó. La prueba de Jung consistía en 100 palabras estímulo. Para cada palabra, el sujeto tenía que responder lo más rápido posible con la primera palabra que le viniera a la mente. Kent y Rosanoff (1910) le dieron al método de asociación un sabor estadounidense distintivo al tabular las reacciones de 1 000 sujetos normales ante una lista de 100 palabras estímulo. Estas tablas se diseñaron para proporcionar una base con la cual comparar las reacciones de sujetos normales y "dementes".

Mientras en EUA se buscaba el enfoque empírico para las pruebas objetivas de personalidad, el joven psiquiatra suizo Hermann Rorschach (1884–1922) desarrollaba un vehículo totalmente diferente para el estudio de la personalidad. Tenía una fuerte influencia del pensamiento jungiano y psicoanalítico, de modo que era natural que su nuevo enfoque se centrara en la tendencia de los pacientes a revelar de manera inconsciente sus conflictos más profundos cuando respondían ante estímulos ambiguos. El Rorschach y otras pruebas proyectivas que se analizan más adelante se fundamentaron en la hipótesis proyectiva: cuando se responde ante estímulos ambiguos o no estructurados, se revelan de modo inadvertido necesidades, fantasías y conflictos más profundos.

Rorschach estaba convencido de que las personas revelaban importantes dimensiones de la personalidad en sus respuestas ante manchas de tinta. Pasó años desarrollando el conjunto

de 10 manchas de tinta y sistemáticamente analizó las respuestas de sus amigos y de diversos grupos de pacientes (Rorschach, 1921). Por desgracia, murió sólo un año después de que se publicó su monografía y a otros les quedó la tarea de terminar su trabajo. Los desarrollos en cuanto al Rorschach se revisan posteriormente dentro del texto.

Mientras la prueba de Rorschach se desarrolló inicialmente para revelar el funcionamiento más profundo del sujeto anormal, el TAT o Test de Apercepción Temática (Morgan y Murray, 1935) se elaboró como un instrumento para estudiar la personalidad normal. Por supuesto, desde entonces ambos se han ampliado para examinar todo el continuo de la conducta humana.

El TAT consiste en una serie de ilustraciones que representan, en su mayor parte, a una o más personas que participan en una interacción ambigua. Al sujeto se le muestra una ilustración a la vez y se le pide que invente una historia acerca de ella. Se le instruye a ser tan riguroso como sea posible, a que discuta los pensamientos y sentimientos y que describa el pasado, presente y futuro de lo que se representa en la lámina.

Murray (1938) consideraba que las necesidades subyacentes de la personalidad, como la necesidad de logro, se revelarían en el contenido de las historias. Aunque se han desarrollado numerosos sistemas de calificación, en general los clínicos han dependido del análisis impresionista para llegar a conclusiones con respecto a los protocolos del TAT. Las aplicaciones modernas del TAT se analizan en un capítulo posterior.

La técnica de frases incompletas también se inició durante esta época con el trabajo de Payne (1928). Se han realizado numerosas extensiones y variaciones de la técnica, que consiste en darle al sujeto una parte de una oración como: "Me aburro cuando _____", y pedirle que la termine. Algunas aplicaciones modernas se analizarán después, pero por el momento se puede mencionar que el problema de la calificación e interpretación, que inquietó a los primeros en desarrollar una prueba de frases incompletas, todavía existe en la actualidad.

Un enfoque totalmente nuevo de las pruebas proyectivas lo emprendió Goodenough (1926), quien trató de determinar no sólo el nivel intelectual, sino también los intereses y rasgos de personalidad de los niños a través del análisis de sus dibujos. La prueba de Buck (1948), llamada Técnica del dibujo proyectivo de Casa-Árbol-Persona, estaba un poco más estandarizada y tenía mayor estructura y requería que el sujeto dibujara una casa, un árbol y una figura humana. El trabajo de Machover (1949) en *Personality Projection in the Drawing of the Human Figure* fue una extensión lógica de los trabajos anteriores. El dibujo de la figura humana como un enfoque proyectivo para comprender la personalidad todavía se emplea; en un capítulo posterior se analizarán los avances modernos en esta práctica.

Mientras tanto, las pruebas proyectivas en Europa estaban dominadas por el Szondi Test, un absurdo instrumento basado en premisas totalmente fallidas. Lipot Szondi fue un psiquiatra suizo de origen húngaro que creía que los principales trastornos psiquiátricos eran causados por genes recesivos. Su prueba constaba de 48 fotografías de pacientes psiquiátricos divididos en seis conjuntos de los siguientes ocho tipos: homosexual, epiléptico, sádico, histérico, catatónico, paranoide, maniaco y depresivo (Deri, 1949). De cada conjunto de ocho fotografías, se le pedía al sujeto que seleccionara aquéllas dos que más le agradaban y las dos que más le desagradaban. Se suponía que una persona que de manera consistente prefiriera un tipo de fotografía entre los seis conjuntos, tendría algunos genes recesivos que le harían sentir simpatía por la persona representada. Así, se suponía que las preferencias proyectivas revelaban genes recesivos que predisponían al individuo a presentar trastornos psiquiátricos específicos.

Deri (1949) importó la prueba a EUA y cambió su fundamentación. Esta investigadora no abogó por una explicación genética recesiva de la elección de fotografías, sino que explicó dichas preferencias con base en la identificación inconsciente con las características de los pacientes fotografiados. Ésta era una base

teórica más admisible para la prueba que las dudosas teorías genéticas de Szondi. Sin embargo, la investigación empírica arrojó dudas sobre la validez de esta prueba, y en poco tiempo este instrumento se quedó en el olvido (Borstel-mann y Klopfer, 1953).

DESARROLLO DE LOS INVENTARIOS DE INTERESES

Mientras los clínicos desarrollaban medidas para analizar la personalidad y los conflictos inconscientes, otros psicólogos diseñaban medidas para la orientación y guía psicológicas de los grandes conjuntos de personas más normales. El inventario de intereses fue la principal de estas medidas, cuyas raíces llegan hasta el estudio de Thorndike (1912) sobre las tendencias del desarrollo de intereses de 100 estudiantes universitarios. En 1919–20, Yoakum desarrolló un conjunto de 1 000 reactivos relacionados con los intereses desde la infancia hasta la madurez temprana (DuBois, 1970). Muchos de estos reactivos se incorporaron en el Inventario Carnegie de Intereses. Cowdery (1926-27) mejoró y refinó el trabajo anterior en el instrumento Carnegie al incrementar el número de reactivos, comparando las respuestas de tres grupos criterio (médicos, ingenieros y abogados) con grupos control de individuos no profesionales y desarrollando una fórmula de ponderación para los reactivos. También fue el primer psicómetra en darse cuenta de la importancia de la validación cruzada. Sometió a prueba sus nuevas escalas en grupos adicionales de médicos, ingenieros y abogados para garantizar que las discriminaciones encontradas en los estudios originales fueran diferencias grupales confiables más que capitalizaciones de la varianza de error.

Edward K. Strong (1884–1963) revisó la prueba de Cowdery y dedicó 36 años al desarrollo de las pautas empíricas para el instrumento modificado como *Strong Vocational Interest Blank* (Formulario Strong de Intereses Vocacionales, SVIB). Puede calificarse a las personas que realizan la prueba con claves separadas para varias docenas de ocupaciones, lo cual proporciona una serie de puntuaciones de inmenso valor para la orientación vocacional. El SVIB se volvió una de las pruebas de más amplio uso de todos los tiempos (Strong, 1927). Su versión moderna, el *Strong Interest Inventory* (Inventario Strong de Intereses), todavía se utiliza de manera general en la orientación vocacional.

Durante decenios, el único competidor serio del SVIB fue el Registro de Preferencias de Kuder (Kuder, 1934), que difería del Strong en cuanto a que forzaba las opciones dentro de tríadas de reactivos. El Kuder era una prueba ipsativa, es decir, comparaba la relativa fortaleza de los intereses dentro del individuo, en vez de comparar las respuestas con diversos grupos profesionales. Las revisiones más recientes del Registro de Preferencias de Kuder incluyen la *Kuder General Interest Survey* (Encuesta de Intereses Generales de Kuder) y la *Kuder Occupational Interest Survey* (Encuesta de Intereses Ocupacionales de Kuder; Kuder, 1966; Kuder y Diamond, 1979; Zytowski, 1985).

RESUMEN DE LOS PRINCIPALES HITOS EN LA HISTORIA DE LAS PRUEBAS

Se concluye esta reseña histórica de las pruebas psicológicas con un resumen en forma de cuadro acerca de los acontecimientos determinantes hasta 1950 (cuadro 1–2). El lector interesado podrá encontrar un listado más detallado —incluyendo cronología de los desarrollos posteriores a 1950— en el Apéndice A.

Cuadro 1–2. Resumen de los primeros hitos en la historia de las pruebas

2200 a.C.	China comenzó con los exámenes para el servicio público
1862	Wilhelm Wundt utiliza un péndulo calibrado para medir la "velocidad de pensamiento"
1884	Francis Galton aplica la primera batería de pruebas a miles de ciudadanos en la Exhibición Internacional de Salud
1890	James McKeen Cattell utiliza el término **prueba mental** al anunciar los puntos esenciales para su batería galtoniana de pruebas
1901	Clark Wissler descubre que los "instrumentos de bronce" de Cattell no se correlacionan con las calificaciones universitarias
1905	Binet y Simon inventan la primera prueba moderna de inteligencia
1914	Stern introduce el CI o cociente de inteligencia: la edad mental dividida entre la edad cronológica
1916	Lewis Terman revisa las escalas Binet-Simon, publica la Stanford-Binet. Aparecen revisiones en 1937, 1960 y 1986
1917	Robert Yerkes encabeza el desarrollo de los exámenes Army Alfa y Army Beta utilizados para evaluar a los reclutas de la primera Guerra Mundial
1917	Robert Woodworth desarrolla la Hoja de Datos Personales, la primera prueba de personalidad
1920	Se publica la prueba Rorschach de Manchas de Tinta
1921	Cattell, Thorndike y Woodworth fundan la *Psychological Corporation*—la primera editorial importante de pruebas
1927	Se publica la primera edición del Estudio de Intereses Vocacionales de Strong
1939	Se publica la Escala Wechsler-Bellevue de Inteligencia
	Revisiones publicadas en 1955, 1981 y 1997
1942	Publicación del Inventario Multifásico de la Personalidad de Minnesota
1949	Se publica la Escala Wechsler de Inteligencia para Niños
	Revisión publicada en 1974, 1991

RESUMEN

1. En 1910, Henry Goddard tradujo la escala Binet-Simon de 1908. En 1911, sometió a prueba a más de mil escolares con base en las normas francesas originales. Le preocupó el hecho de encontrar que 3% de la muestra fueran "débiles mentales" y recomendó que estos niños se aislaran del resto de la sociedad.
2. Las pruebas no verbales de inteligencia se inventaron a inicios del siglo XX para facilitar el examen de los inmigrantes que no eran de habla inglesa. Por ejemplo, Knox publicó en 1914 una prueba con rompecabezas de madera y también utilizó la ahora familiar prueba de sustitución de dígitos por símbolos.
3. En 1916, Lewis Terman publicó la Stanford-Binet, una revisión de las escalas Binet. Esta prueba bien diseñada y con normas cuidadosas, colocó de una vez por todas al examen de la inteligencia sobre una base firme.
4. Durante la primera Guerra Mundial, Robert Yerkes encabezó un equipo de psicólogos que produjeron la prueba Army Alfa, una

prueba grupal con carga verbal para reclutas promedio y superiores, y la Army Beta, una prueba grupal no verbal dirigida a reclutas analfabetos o no angloparlantes.

5. Los pioneros en pruebas, como C. C. Brigham, utilizaron los resultados de las pruebas individuales y grupales para fundamentar las diferencias étnicas en inteligencia y, por tanto, para justificar las restricciones a la inmigración. Posteriormente, algunos de estos pioneros se retractaron de sus puntos de vista anteriores.
6. Las pruebas educativas cayeron dentro del ámbito del Comité de Examen de Ingreso a la Universidad (CEEB), fundado a principios del siglo XX. En 1947, el CEEB fue reemplazado por el Servicio de Exámenes Educativos (ETS), que supervisó la publicación de pruebas tan conocidas como la Prueba de Aptitudes Escolares y el Examen de Selección para el Posgrado.
7. El advenimiento de las baterías de prueba de aptitudes múltiples se hizo posible debido al desarrollo del análisis factorial por parte de L. L. Thurstone y otros. Posteriormente, la mejoría de estas baterías de prueba se vio alentada por la necesidad práctica de seleccionar a los reclutas de la Segunda Guerra Mundial para puestos de alta especialización.
8. Las pruebas de personalidad comenzaron con la *Woodworth's Personal Data Sheet*, una simple lista de verificación de síntomas con respuestas de sí y no que se utilizó para detectar psiconeurosis en los reclutas de la segunda Guerra Mundial. Muchos inventarios posteriores, incluyendo al popular Inventario Multifásico de la Personalidad de Minnesota, tomaron prestado el contenido de la hoja de datos personales.
9. Las pruebas proyectivas comenzaron con la técnica de asociación de palabras, cuyo pionero fue Francis Galton y que concretó C. G. Jung en 1910. Hermann Rorschach publicó su famosa prueba de manchas de tinta en 1921.
10. El Test de Apercepción Temática (TAT), una prueba de narración de historias ante ilustraciones, introducida en 1935 por Morgan y Murray, se basó en la hipótesis proyectiva: cuando responden a estímulos ambiguos o no estructurados, los examinados revelan de manera inadvertida sus necesidades, fantasías y conflictos más profundos.
11. La evaluación de los intereses vocacionales comenzó con el *Carnegie Interest Inventory* (Inventario Carnegie de Intereses), de Yoakum, que se desarrolló en 1919–20. Después de varias revisiones y extensiones, este instrumento surgió como el *Strong's Vocational Interest Blank* (SVIB), de E. K. Strong. ■

CAPÍTULO 2

Pruebas y su proceso de aplicación

TEMA 2A
Naturaleza y usos de las pruebas psicológicas

Definición de una prueba
Otras distinciones en pruebas
Tipos de pruebas
Usos de las pruebas
Quién puede obtener las pruebas
Fuentes de información sobre pruebas
Resumen
Términos y conceptos clave

La introducción histórica en el capítulo anterior sólo ha familiarizado al lector con una pequeña fracción de los muchos tipos de pruebas psicológicas y sus usos. Las primeras se utilizaron de manera predominante para dos propósitos: medir la inteligencia y detectar trastornos de personalidad. Por tanto, es comprensible que el ciudadano promedio haga una equivalencia entre pruebas psicológicas y puntuaciones de CI, manchas de tinta e inventarios de personalidad. Con toda seguridad, en este punto de vista existe más que un grano de verdad: las medidas de personalidad e inteligencia todavía son los pilares esenciales de las pruebas psicológicas. Sin embargo, los psicómetras han desarrollado muchos otros tipos de estos instrumentos para propósitos diversos e imaginativos que los pioneros nunca podrían haber anticipado. Este capítulo proporciona una visión panorámica de las pruebas psicológicas y de sus numerosas aplicaciones. En el tema 2A, Naturaleza y usos de las pruebas psicológicas, se resumen los diferentes tipos y las diversas aplicaciones de éstas. En el tema 2B, Proceso

de Aplicación, se enfatiza que la aplicación de pruebas es una transacción entre el examinador y la persona evaluada, no un proceso estéril de medición.

Desde el nacimiento hasta la vejez, nos topamos con estos instrumentos en casi todos los momentos de cambio en la vida. La primera prueba del bebé, realizada inmediatamente después del nacimiento, es la prueba Apgar, una evaluación rápida y multivariada de la frecuencia cardiaca, respiración, tono muscular, irritabilidad refleja y color (Clarke-Stewart y Friedman, 1987). La puntuación total Apgar (0 a 10) ayuda a determinar la necesidad de cualquier tipo de atención médica inmediata. Después, un infante que ha recibido antes una baja puntuación Apgar podría ser candidato para una evaluación de discapacidades del desarrollo. El niño preescolar puede realizar pruebas de preparación para la escuela. Una vez que se ha comenzado con la carrera escolar, cada estudiante atraviesa por cientos, quizá miles, de pruebas académicas antes de graduarse (sin mencionar las que miden discapacidades para el aprendizaje, inteligencia dotada, intereses vocacionales y admisión a la universidad). Después de graduarse, los adultos pueden enfrentar pruebas para ingresar a un empleo, obtener la licencia de manejo, autorización de seguridad, funcionamiento de personalidad, compatibilidad matrimonial, discapacidades del desarrollo, disfunción cerebral —la lista es casi interminable. Algunas personas incluso se enfrentan a una última indignidad en la parte más débil de sus últimos años: una prueba para determinar su capacidad para administrar sus asuntos económicos.

La idea de una prueba es, por tanto, un elemento que domina en nuestra cultura, una característica que damos por sentada. Sin embargo, el concepto que tiene un lego acerca de ellas no necesariamente coincide con la perspectiva más restringida que tiene un **psicómetra** (especialista en psicología o educación que desarrolla y evalúa pruebas psicológicas). Debido a los equívocos generales en cuanto a la comprensión de su naturaleza, es adecuado comenzar este tema con una pregunta fundamental que define el campo de acción de todo el libro: ¿qué es una prueba?

DEFINICIÓN DE UNA PRUEBA

Una **prueba** es un procedimiento estandarizado para tomar una muestra de conducta y describirla con categorías o puntuaciones. Además, la mayor parte tiene normas o estándares, con base en los cuales pueden utilizarse los resultados para pronosticar otras conductas más importantes. En las siguientes secciones se hablará más sobre estas características, pero primero es útil representar el campo de acción de la definición. Incluidas dentro de esta perspectiva se encuentran las pruebas tradicionales como los cuestionarios de personalidad y las pruebas de inteligencia, pero la definición también incluye diversos procedimientos que el lector podría no reconocer como tales. Por ejemplo, todos los métodos siguientes pueden ser pruebas, según la definición que se utiliza en este libro: una lista de verificación para clasificar las habilidades sociales de un joven con retraso mental; una medida sin límite de tiempo del dominio en la suma de pares de números de tres dígitos; valoraciones por computadora del tiempo de reacción; e incluso, pruebas situacionales como la observación de un individuo que trabaja en una tarea de grupo con dos "ayudantes" poco cooperativos y que sólo le causan dificultades.

En suma, las pruebas son sumamente variadas en sus formatos y aplicaciones. Sin embargo, la mayor parte posee estas características que las definen:

- Procedimiento estandarizado.
- Muestra de conducta.
- Puntuaciones o categorías.
- Normas o estándares.
- Predicción de conducta fuera de la prueba.

En las siguientes secciones se analizarán con más detalle cada una de estas características. La imagen que se desea representar tiene que ver en especial con las pruebas referentes a la norma —las que utilizan una población bien definida de personas para su esquema interpretativo. Sin embargo, las características definitorias de una prueba difieren un tanto en

el caso especial de las pruebas referidas a criterio —las que miden lo que una persona puede hacer, en lugar de comparar los resultados con los niveles de desempeño de otros. Por tal razón, se tratan por separado las pruebas referidas a criterio.

El **procedimiento estandarizado** es una característica esencial de cualquier prueba psicológica. Se considera que una prueba está **estandarizada** si los procedimientos para su aplicación son uniformes de un examinador a otro y de un ambiente a otro. Por supuesto, la estandarización depende, en cierto grado, de la capacidad del examinador. Incluso la mejor prueba puede resultar inútil en manos de un evaluador descuidado, con entrenamiento deficiente o mal informado, como descubrirá el lector en el tema 2B, Proceso de aplicación. Sin embargo, la mayoría de los examinadores son competentes. Por tanto, la estandarización depende en gran medida de las disposiciones de aplicación que se encuentran en el manual de instrucciones que por lo común acompaña a una prueba.

La formulación de las instrucciones es un paso esencial para la estandarización de una prueba. A fin de garantizar procedimientos uniformes de aplicación, quien desarrolla una prueba debe proporcionar materiales estímulo comparables para todos los examinados, debe especificar con una precisión considerable las instrucciones verbales para cada reactivo o subprueba y debe aconsejarle al examinador cómo ha de manejar una diversidad de dudas por parte de la persona evaluada.

Para ilustrar estos puntos, considérense las diversas maneras en que una persona que desarrolla una prueba podría enfocar la evaluación de la **retención de dígitos** —el número máximo de dígitos presentados verbalmente que un sujeto puede recordar de memoria. Podría ser que una prueba no estandarizada de retención de dígitos sugiriera tan sólo que el examinador presentara de manera verbal series cada vez más largas de números hasta que el sujeto fallara. El número de dígitos en la serie más larga recordada sería, entonces, la capacidad de retención de dígitos del sujeto. La mayoría de los lectores puede darse cuenta de que una prueba con tal definición tan general carecerá de uniformidad de un examinador a otro. Si quien aplica la prueba está en libertad de improvisar cualquier serie de dígitos, ¿qué podría impedirle que presentara, con la inflexión familiar de un locutor de televisión, "1-800-325-3535"? Tal serie sería bastante más fácil de recordar que un conjunto más aleatorio, por ejemplo, "7-2-8-1-9-4-6-3-7-4-2". La velocidad de presentación también puede tener un efecto crucial sobre la uniformidad de una prueba de retención de dígitos. Para propósitos de estandarización, es esencial que todos los examinadores presenten cada serie a una tasa constante; por ejemplo, un dígito por segundo. Por último, el examinador necesita saber cómo ha de reaccionar ante respuestas inesperadas, como el que un sujeto le diga: "¿podría repetirlos de nuevo?" Por razones obvias, el consejo habitual es "no".

Quien desarrolla una prueba puede incluso llegar hasta el punto de recomendar el comportamiento deseado en el examinador, como mantener una expresión facial neutra cuando se registra la respuesta de un sujeto. Estas influencias aparentemente sutiles pueden tener un serio impacto sobre la uniformidad de los procedimientos de prueba. Por ejemplo, un examinador que se sonríe con displicencia cuando registra las respuestas podría provocar que el sujeto se sienta ansioso y falle en una tarea fácil. En el siguiente tema, el Proceso de aplicación, se analizará la influencia potencial del examinador sobre los resultados de prueba.

Una prueba psicológica también es una muestra limitada de conducta. Ni el sujeto ni el examinador tienen el tiempo suficiente para una prueba realmente amplia, incluso cuando ésta se dirige a un dominio conductual bien definido y finito. Así, las restricciones prácticas dictan que una prueba es sólo una muestra de conducta. Sin embargo, la muestra de conducta es de interés sólo en la medida que permita que el examinador realice inferencias acerca del dominio total de conductas relacionadas. Por ejemplo, el propósito de una prueba de vocabulario consiste en determinar la reserva completa

de palabras de la persona examinada, al pedir definiciones de una muestra muy pequeña, pero cuidadosamente seleccionada, de palabras. El hecho de que el sujeto pueda definir las 35 palabras particulares de una subprueba de vocabulario (p. ej., en el WAIS-R) es de poca consecuencia directa. Pero el significado de dichos resultados es de gran importancia debido a que señala el conocimiento general de vocabulario del individuo evaluado.

Un punto interesante —del que el público en general tiene poco conocimiento— es que los reactivos de prueba no necesitan parecerse a las conductas que la prueba intenta pronosticar. La característica esencial de una buena prueba es que permita que el examinador pronostique otras conductas —no que refleje aquéllas a predecir. Si responder "cierto" a la afirmación "bebo mucha agua" ayuda a predecir la depresión, entonces esta afirmación, en apariencia no relacionada, es un índice útil de la depresión. Por tanto, el lector observará que la predicción exitosa es una cuestión empírica que se resuelve a través de la investigación apropiada. Aunque la mayoría de las pruebas toman una muestra directa del dominio de conductas que esperan predecir, éste no es un requisito psicométrico.

Una prueba psicológica también debe permitir la derivación de puntuaciones o categorías. Thorndike (1918) expresó el axioma esencial de las pruebas en su famosa aseveración de que: "aquello que existe de alguna manera, existe en cierta cantidad". McCall (1939) fue un paso más allá al declarar: "cualquier cosa que existe en cierta cantidad, puede medirse". Las pruebas se esfuerzan en ser una forma de medición similar a los procedimientos de las ciencias físicas, donde los números representan dimensiones abstractas como peso o temperatura. Toda prueba suministra una o más puntuaciones o proporciona evidencia de que una persona pertenece a una categoría y no a otra. En pocas palabras, las pruebas psicológicas resumen la ejecución en números o clasificaciones.

La suposición implícita del punto de vista psicométrico es que las pruebas miden las diferencias individuales en cuanto a rasgos o características que existen en cierto sentido vago de la palabra. En la mayor parte de los casos, se supone que todas las personas poseen el rasgo o característica medida, aunque en diferentes cantidades. El propósito de la prueba consiste en estimar la cantidad del rasgo o cualidad que posee un individuo.

En este contexto, se deben hacer dos advertencias. Primero, toda puntuación de prueba siempre reflejará cierto grado de error de medición. La imprecisión de las pruebas es simplemente inevitable: éstas deben depender de una muestra externa de conducta para estimar la característica no observable y, por tanto, inferida. Con frecuencia, los psicómetras expresan este punto fundamental con la ecuación:

$$X = T + e$$

donde X es la puntuación observada, T la puntuación verdadera y e el componente de error positivo o negativo. Lo mejor que puede hacer quien desarrolla una prueba es procurar que e sea muy pequeño. Nunca se le puede eliminar por completo, como tampoco se puede saber su impacto exacto sobre el caso individual. El concepto de error de medición se analizará en el tema 3B, Conceptos de confiabilidad.

La segunda advertencia es que los consumidores de pruebas deben prevenirse de materializar las características medidas. Los resultados no representan una "cosa" que tenga realidad física; por lo común, representan una abstracción que se ha demostrado que tiene utilidad para pronosticar conductas externas a la prueba. Por ejemplo, cuando se analiza el CI de una persona, los psicólogos se refieren a una abstracción que no tiene existencia directa, material, pero que, sin embargo, es útil para pronosticar el rendimiento escolar y otros resultados.

Una prueba psicológica también debe poseer normas o estándares. En general, la puntuación de prueba de una persona se interpreta al compararla con las puntuaciones obtenidas por otros individuos en la misma prueba. Para este propósito, es común que quienes desarrollan

pruebas aplican **normas** —un resumen de los resultados de prueba de un grupo grande y representativo de personas (Petersen, Kolen y Hoover, 1989). El grupo normativo se conoce como muestra de estandarización.

La selección y evaluación de la **muestra de estandarización** es crucial para la utilidad de una prueba. Este grupo debe ser representativo de la población a la que se dirige la prueba o, de otra manera, no será posible determinar la posición relativa de un individuo examinado. En el caso extremo donde no se proporcionan normas, el examinador no puede utilizar en absoluto los resultados de la prueba. Una excepción a este punto ocurre en el caso de las pruebas referidas a criterio, las cuales se analizan más adelante.

Las normas no sólo establecen un desempeño promedio, sino que también sirven para indicar la frecuencia con la que se obtienen diferentes puntuaciones altas y bajas. Así, las normas permiten que el examinador determine el grado al que se desvía una puntuación con respecto a las expectativas. Tal información puede ser muy importante en la predicción de la conducta externa a la prueba del individuo examinado. Las normas tienen tal importancia trascendental en la interpretación de pruebas, que después se les considerará con amplitud en una sección independiente dentro de este texto.

Por último, las pruebas no constituyen un fin en sí mismas. En general, el propósito último de una prueba consiste en pronosticar conductas adicionales, diferentes de aquellas que se muestrean de manera directa en la prueba. Así, el examinador puede tener mayor interés en las conductas externas a la prueba pronosticadas por ésta que en las respuestas de prueba en sí. Quizás un ejemplo concreto aclarará este punto. Supóngase que un examinador aplica una prueba de manchas de tinta a un paciente en un hospital psiquiátrico y que el paciente responde a una mancha de tinta describiéndola como "ojos que se asoman sigilosamente". Con base en las normas establecidas, el examinador podría pronosticar entonces que el sujeto es sumamente suspicaz y obtendrá poco beneficio de una psicoterapia individual. El propósito de la prueba consiste en llegar a ésta y otras predicciones similares —no determinar si la persona percibe ojos que lo miran fijamente desde las manchas.

La capacidad de una prueba para pronosticar conducta externa se determina por un amplio cuerpo de investigación de validación, cuya mayoría se realiza después de que se ha publicado la prueba, pero no hay ninguna garantía en el mundo de la investigación psicométrica. Es común que un investigador publique una prueba prometedora, sólo para leer años después que otros investigadores la han encontrado deficiente. Ésta es una lección para quienes emplean las pruebas: el hecho de que una prueba exista y declare medir cierta característica, no es garantía de que sea veraz en sus afirmaciones. Una prueba puede tener un bonito título, instrucciones precisas, normas elaboradas, empaque atractivo y descubrimientos preliminares —pero si en el estudio desapasionado por parte de investigadores independientes, la prueba no puede pronosticar conductas externas apropiadas, entonces es inútil.

OTRAS DISTINCIONES EN PRUEBAS

Las principales características de una prueba, que se resumieron antes, se aplican en particular a las referidas a la norma, que constituyen gran parte de las pruebas en uso. En una **prueba referida a la norma**, el desempeño de cada examinado se interpreta con referencia a una muestra de estandarización relevante (Petersen, Kolen y Hoover, 1989). Sin embargo, estas características son menos importantes en el caso especial de la **prueba referida a criterio**, dado que estos instrumentos no necesitan de la comparación del individuo en particular con un grupo de referencia. En este tipo de instrumento, el objetivo consiste en determinar la posición de la persona evaluada con respecto a objetivos educativos definidos de manera muy estrecha (Berk, 1984). Por ejemplo, una parte de una prueba de aritmética para niños de 10 años de

edad podría medir el nivel de precisión en la suma de pares de números de dos dígitos. En una prueba sin límite de tiempo con 20 de dichos problemas, la precisión sería casi perfecta. Para este tipo de prueba, en realidad no importa cómo se compara al individuo examinado con otros de la misma edad, lo que importa es si el individuo satisface un criterio apropiado y específico —por ejemplo, precisión de 95%. Debido a que no existe comparación con el desempeño normativo de otros, este tipo de herramienta de medición se ha llamado correctamente prueba referida a criterio, la cual, a diferencia de las pruebas referidas a la norma, puede interpretarse de manera significativa sin hacer referencia a normas. En el tema 3A, Normas y estandarización, se analizarán con mayor detalle estas pruebas.

Otra distinción importante existe entre los términos prueba y evaluación, que con frecuencia se consideran equivalentes, sin embargo, no quieren decir exactamente lo mismo. **Evaluación** es un término más amplio, que se refiere a todo el proceso de recopilar información sobre una persona y utilizarla para pronosticar la conducta. Puede definirse como la valoración o estimación de la magnitud de uno o más atributos en una persona. La evaluación de las características humanas implica observaciones, entrevistas, listas de verificación, inventarios, pruebas proyectivas y otras pruebas psicológicas. En suma, las pruebas representan sólo una fuente de información utilizada en el proceso de evaluación, en el cual el examinador debe comparar y combinar los datos de diferentes fuentes. Éste es un proceso subjetivo de manera inherente, que requiere que el examinador seleccione entre la información conflictiva y realice predicciones con base en la gestalt completa de datos.

El término **evaluación** se inventó durante la segunda Guerra Mundial para describir un programa dirigido a seleccionar hombres que destacaran en el servicio secreto, dentro de la *Office of Strategic Services* (OSS Assessment Staff, 1948). El personal de psicólogos y psiquiatras de la OSS amasaba una inmensa cantidad de información acerca de los candidatos durante cuatro agotadores días de pruebas escritas, entrevistas y pruebas de personalidad. Además, el proceso de evaluación incluía una variedad de pruebas sobre situaciones de la vida real que se basaban en la conciencia de que existe una diferencia entre saber hacer y poder hacer:

> ... hicimos que los candidatos intentarán, de hecho, las tareas de manera física o verbal, en vez de tan sólo indicar por escrito cómo podrían hacerse. Se nos impulsó a introducir pruebas realistas de la capacidad debido a descubrimientos como el siguiente: los hombres que obtienen una alta puntuación en Comprensión Mecánica, una prueba de lápiz y papel, pueden encontrarse por debajo del promedio cuando se trata de resolver problemas mecánicos con sus manos (OSS Assessment Staff, 1948).

Las pruebas situacionales incluían tareas de grupo para transportar equipo al otro lado de un arroyo y escalar una pared de 3 m de altura, así como el escrutinio individual de la capacidad para sobrevivir un interrogatorio realista y comandar a dos subordinados poco cooperativos en una tarea de construcción.

Con base en las observaciones conductuales y en los resultados de prueba, el personal de la OSS calificaba a los candidatos en docenas de rasgos específicos, en categorías tan amplias como liderazgo, relaciones sociales, estabilidad emocional, inteligencia efectiva y capacidad física. Estas calificaciones sirvieron para seleccionar al personal militar de la OSS.

TIPOS DE PRUEBAS

Las pruebas pueden agruparse de manera amplia en dos bandos: pruebas grupales y pruebas individuales. Las **pruebas grupales** son medidas principalmente de lápiz y papel, adecuadas para examinar a grandes grupos de personas a un mismo tiempo. Las **pruebas individuales** son instrumentos que, por su diseño y propósito, deben aplicarse a una sola persona. Una ventaja importante de estas últimas es que el examinador puede estimar el nivel de motivación

del sujeto y evaluar la relevancia de otros factores (p. ej., impulsividad o ansiedad) sobre los resultados de prueba.

Por conveniencia, las pruebas se clasificarán en ocho categorías representadas en el cuadro 2–1. Cada una de ellas contiene pruebas referidas a la norma, referidas a criterio, individuales y grupales. El lector observará que cualquier tipología de las pruebas es una determinación puramente arbitraria. Por ejemplo, podría postularse incluso otra dicotomía: pruebas que buscan medir el desempeño máximo (p. ej., una prueba de inteligencia) contra las que buscan estimar una respuesta típica (p. ej., un inventario de personalidad).

En un sentido estricto, existen cientos —quizá miles— de tipos diferentes de pruebas, cada una de las cuales mide un aspecto ligeramente diferente del individuo. Por ejemplo, podría discutirse que incluso dos pruebas de inteligencia constituirían diferentes tipos de medida. Una prueba podría revelar la suposición de que la inteligencia es un constructo biológico que puede medirse mejor a través de las ondas cerebrales, mientras otra podría fundamentarse en la perspectiva tradicional de que la inteligencia se exhibe en la capacidad para aprender habilidades aculturadas como el vocabulario. Agrupar ambas medidas bajo la categoría de **pruebas de inteligencia** es con toda seguridad una simplificación exagerada pero, sin embargo, es un punto de partida útil.

Como se vio en el primer capítulo, las **pruebas de inteligencia** se diseñaron originalmente para tomar una muestra de un amplia variedad de habilidades, a fin de estimar el nivel intelectual general del individuo. Las escalas Binet-Simon tuvieron éxito, en parte, debido a que incorporaban tareas heterogéneas, incluyendo definiciones de palabras, memoria de diseños, preguntas de comprensión y tareas de visualización espacial. Las pruebas grupales de inteligencia que florecieron con tal profusión durante y después de la segunda Guerra Mundial también medían capacidades diversas —como lo demuestra la prueba Army Alfa con sus ocho secciones diferentes que miden juicio práctico, información, aritmética y razonamiento, entre otras habilidades.

Las pruebas modernas de inteligencia también emulan este patrón históricamente establecido al tomar una muestra de una amplia

Cuadro 2–1. Principales tipos de pruebas psicológicas

Pruebas de inteligencia: miden la capacidad de un individuo en áreas relativamente globales como comprensión verbal, organización perceptual o razonamiento y, por tanto, ayudan a determinar el potencial para el trabajo escolar o para ciertas ocupaciones.

Pruebas de aptitud: miden la capacidad para una tarea o tipo de habilidad relativamente específicas; en efecto, las pruebas de aptitud son una forma estrecha de prueba de capacidad.

Pruebas de aprovechamiento: miden el grado de aprendizaje, éxito o logro de una persona en un tema o tarea.

Pruebas de creatividad: evalúan el pensamiento novedoso y original y la capacidad para encontrar soluciones poco comunes o inesperadas, en especial, para problemas definidos de manera vaga.

Pruebas de personalidad: miden los rasgos, cualidades o conductas que determinan la individualidad de una persona; dichas pruebas incluyen listas de verificación, inventarios y técnicas proyectivas.

Inventarios de intereses: miden la preferencia de un individuo por ciertas actividades o temas y, por tanto, ayudan a determinar la elección de carrera.

Procedimientos conductuales: describen de manera objetiva una conducta y estiman su frecuencia, identificando los antecedentes y consecuencias de la conducta.

Pruebas neuropsicológicas: miden el desempeño cognoscitivo, sensorial, perceptual y motor para determinar el grado, localización y consecuencias conductuales del daño cerebral.

variedad de destrezas consideradas importantes en nuestra cultura. En general, el término **prueba de inteligencia** se refiere a una prueba que produce una puntuación resumida general, basada en los resultados de una muestra heterogénea de reactivos. Por supuesto, una prueba de este tipo podría también proporcionar un perfil de las puntuaciones de subprueba, pero es la puntuación general la que en términos generales atrae la mayor atención.

Una **prueba de aptitud** mide uno o más segmentos claramente definidos y relativamente homogéneos de una capacidad; tiene dos variedades: pruebas de una sola aptitud y baterías de prueba de aptitudes múltiples. Como es obvio, las primeras evalúan sólo una capacidad, mientras que las segundas proporcionan un perfil de puntuaciones para varias aptitudes.

Con frecuencia, las pruebas de aptitud se emplean para pronosticar el éxito en una profesión, curso de entrenamiento o esfuerzos educativos. Por ejemplo, las Medidas Seashore de Talentos Musicales (Seashore, 1938), una serie de pruebas que cubren tono, sonoridad, ritmo, tiempo, timbre y memoria tonal, pueden utilizarse para identificar a niños que tienen un talento potencial para la música. También existen pruebas de aptitudes especializadas para la evaluación de habilidades en trabajo de oficina, capacidades mecánicas, destreza manual y capacidad artística. Éstas se revisan en el tema 8A, Pruebas de aptitud y análisis factorial.

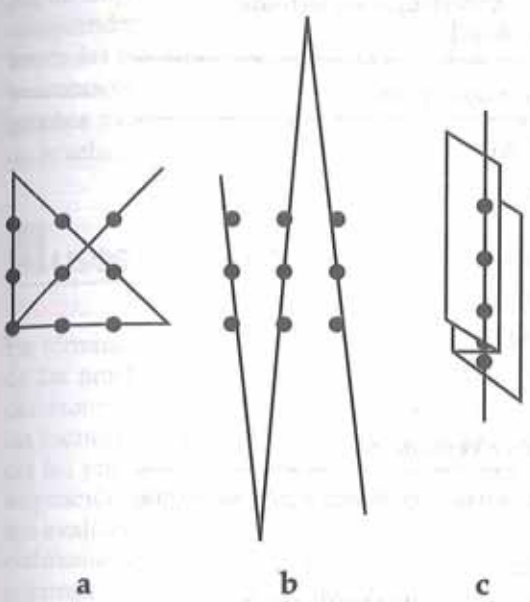
El uso más común para las pruebas de aptitud consiste en determinar las admisiones a la universidad. La mayoría de los estudiantes universitarios están familiarizados con el SAT (del inglés *Scholastic Assessment Test*; Prueba de Evaluación Escolar, antes llamada *Scholastic Aptitude Test*; Prueba de Aptitud Escolar) del Comité de Examen de Ingreso a la Universidad. Esta prueba contiene una sección Verbal, que destaca el conocimiento de palabras y la comprensión de lectura, y una sección de Matemáticas, que destaca el álgebra, la geometría y el razonamiento perspicaz. En efecto, las universidades que requieren con fines de admisión ciertas puntuaciones mínimas en el SAT, utilizan la prueba para pronosticar el éxito académico.

Las **pruebas de aprovechamiento** miden el grado de aprendizaje, éxito o logro de una persona en una materia. La suposición implícita de la mayor parte de estas pruebas es que las escuelas han enseñado la materia de manera directa. Por tanto, el propósito de la prueba consiste en determinar la cantidad del material que el sujeto ha absorbido o dominado. En general, las pruebas de aprovechamiento tienen varias subpruebas; por ejemplo, lectura, matemáticas, lenguaje, ciencias naturales y ciencias sociales, las cuales se revisan en el tema 8B, Pruebas grupales de aprovechamiento.

La distinción entre pruebas de aptitud y de aprovechamiento es más una cuestión de uso que de contenido (Gregory, 1994a). De hecho, cualquier prueba puede ser de aptitud, en el sentido de que ayuda a pronosticar el desempeño futuro. De la misma manera, cualquier prueba puede ser de aprovechamiento, en el sentido de que refleja cuánto ha aprendido el sujeto. Por tanto, en la práctica, la distinción entre estos dos tipos de instrumentos se determina por sus respectivos usos. En ciertas ocasiones, un instrumento puede servir para ambos propósitos, actuando como una prueba de aptitud para predecir el desempeño futuro, y como prueba de aprovechamiento, para supervisar el aprendizaje pasado.

Las **pruebas de creatividad** evalúan la capacidad del sujeto para producir nuevas ideas, discernimientos o creaciones artísticas que se consideran de valor social, estético o científico. Así, las medidas de **creatividad** enfatizan la novedad y originalidad en la solución de problemas confusos o en la producción de obras artísticas. En la figura 2–1 se ilustra una respuesta creativa a un problema.

Las pruebas de creatividad tienen una historia accidentada. En el decenio de 1960 se les promocionaba como una alternativa útil para las pruebas de inteligencia y se les utilizó ampliamente en los sistemas educativos de EUA. Los educadores se mostraron especialmente impresionados de que las pruebas de creatividad requirieran del pensamiento divergente —empleando una variedad de respuestas para un problema complejo o confuso— en oposición



Nota: Sin levantar el lápiz, cruce todos los puntos con el menor número posible de líneas rectas. La solución habitual se muestra en a. En b y c se presentan soluciones creativas.

Figura 2–1. Soluciones para el problema de nueve puntos como ejemplos de creatividad.

al pensamiento convergente —encontrando una solución correcta única para un problema bien definido. Por ejemplo, una prueba de creatividad podría pedirle al individuo examinado que imaginara todas las cosas que sucederían si las nubes tuvieran cuerdas que colgaran de ellas hasta el piso (Guilford, 1954). Se suponía que los alumnos que pudieran dar un gran número de consecuencias eran más creativos que sus compañeros menos imaginativos. Sin embargo, algunos psicómetras se mostraron escépticos, concluyendo que la creatividad es sólo otra etiqueta para la inteligencia aplicada (p. ej., McNemar, 1964).

Las **pruebas de personalidad** miden los rasgos, cualidades o conductas que determinan la individualidad de una persona; esta información ayuda a pronosticar la conducta. Estas pruebas pueden encontrarse en muchas variedades diferentes, incluyendo listas de verificación, inventarios y técnicas proyectivas como frases incompletas y manchas de tinta (cuadro 2–2).

Los **inventarios de intereses** miden la preferencia de un individuo por ciertas actividades o temas y, con ello, ayudan a determinar la elección de carrera; se basan en la suposición explícita de que los patrones de interés determinan y, por tanto, también predicen la satisfacción con el trabajo. Por ejemplo, si la persona examinada tiene los mismos intereses que los contadores exitosos y satisfechos, se considera probable que disfrutará del trabajo de un contador. La suposición de que los patrones de interés pronostican la satisfacción con el trabajo se confirma en gran medida a través de estudios empíricos, como se revisará en el tema 12A, Evaluación de intereses y valores laborales.

Existen muchos tipos de **procedimientos conductuales** para evaluar los antecedentes y consecuencias de la conducta, incluyendo listas de verificación, escalas de clasificación, entrevistas y observaciones estructuradas. Estos métodos comparten una suposición común de que la conducta puede comprenderse mejor en términos de características definidas de manera clara como frecuencia, duración, antecedentes y consecuencias. Los procedimientos conductuales tienden a ser sumamente pragmáticos en el sentido de que generalmente se encuentran entretejidos con enfoques de tratamiento.

Las **pruebas neuropsicológicas** se utilizan para la evaluación de personas de las que se sospecha o se sabe que tienen una disfunción cerebral. La **neuropsicología** es el estudio de las relaciones cerebro-conducta. A través de los años, los neuropsicólogos han descubierto que ciertas pruebas y procedimientos son muy sensibles a los efectos del daño cerebral, y emplean estas pruebas y procedimientos especializados para hacer inferencias acerca de la localización, grado y consecuencias de dicho daño.

Aunque las pruebas y procedimientos neuropsicológicos son útiles para llegar a un diagnóstico neurológico, su principal propósito consiste en valorar las fortalezas y debilidades sensoriales, motoras, cognitivas y conductuales

Cuadro 2–2. Ejemplos de reactivos en una prueba de personalidad

(a) Lista de verificación de adjetivos:

Marque aquellas palabras que lo describen:

() relajado	() asertivo
() atento	() curioso
() alegre	() ecuánime
() impaciente	() escéptico
() taciturno	() impulsivo
() optimista	() ansioso

(b) Inventario de Falso-Verdadero:

Encierre en un círculo la verdad o falsedad de cada afirmación, según se aplique a usted:

V F Me gustan las revistas de deportes.
V F La mayoría de las personas mentirían para obtener un trabajo.
V F Me gustan las grandes fiestas donde hay mucha diversión.
V F Pensamientos extraños me poseen durante horas a la vez.
V F Con frecuencia me arrepiento de las oportunidades que perdí en la vida.
V F En ocasiones me siento ansioso sin razón alguna.
V F Me agradan todas las personas que he conocido.
V F Quedarme dormido es un problema que me ocurre en raras ocasiones.

(c) Prueba proyectiva de frases incompletas:

Termine cada oración con el primer pensamiento que le venga a la mente:

Me siento aburrido cuando.
Lo que más necesito es.
Me agradan las personas que.
Mi madre era.

de un paciente con alteración neurológica.* La evaluación de fortalezas y debilidades en estos pacientes es crucial para documentar la mejoría, proyectar el grado de deterioro en los padecimientos degenerativos y planificar tratamientos efectivos para incapacidades específicas. Una evaluación neurológica completa requiere por lo común de 3 a 8 horas de pruebas individuales con una amplia batería de medidas. Los examinadores deben haber atravesado

* Las técnicas radiológicas avanzadas como la Tomografía Computarizada (TC), las Imágenes por Resonancia Magnética (IRM) y la Tomografía con Emisión de Positrones (TEP) permiten ahora que los neurólogos realicen inferencias sumamente precisas acerca de la presencia, localización y causas del daño cerebral. Sin embargo, esto no reduce la importancia de las pruebas neuropsicológicas en la determinación de las **consecuencias funcionales** del daño cerebral en la vida del paciente individual.

por un amplio entrenamiento avanzado, a fin de comprender la gran cantidad de datos resultantes de las pruebas. En el tema 9B, Evaluación neuropsicológica y geriátrica, se revisarán las pruebas individuales y las principales baterías de prueba.

USOS DE LAS PRUEBAS

En términos generales, el empleo más común de las pruebas psicológicas consiste en tomar decisiones acerca de las personas. Por ejemplo, las instituciones educativas utilizan con frecuencia las pruebas para determinar los niveles de asignación para los alumnos y las universidades evalúan a quién deben admitir, basadas parcialmente en las puntuaciones de prueba. Los sistemas del servicio público estatales, federales y locales también dependen, en gran medida, de las pruebas para propósitos de selección de personal.

Incluso el profesional independiente utiliza principalmente las pruebas para la toma de decisiones. Los ejemplos incluyen al psicólogo consultor que emplea una prueba de personalidad para determinar si un departamento de policía contrata a un candidato y no a otro y al neuropsicólogo que emplea pruebas para concluir que un cliente ha sufrido daño cerebral.

Pero la simple toma de decisiones no es la única función de las pruebas psicológicas. Es conveniente distinguir cinco usos de las pruebas:

- Clasificación.
- Diagnóstico y planificación del tratamiento.
- Autoconocimiento.
- Valoración de programas.
- Investigación.

Con frecuencia, estas aplicaciones se traslapan y, en ciertas ocasiones, es difícil distinguir unas de otras. Por ejemplo, una prueba que ayuda a determinar un diagnóstico psiquiátrico podría también proporcionar una forma de autoconocimiento. Se analizarán con mayor detalle estas aplicaciones.

El término **clasificación** engloba una variedad de procedimientos que comparten un propósito común: asignar a una persona a una categoría en vez de a otra. Por supuesto, la asignación de categorías no es un fin en sí mismo, sino la base para un tratamiento diferencial de algún tipo. Así, la clasificación puede tener importantes efectos, como conceder o restringir el acceso a una universidad específica o determinar si se contrata a una persona para un trabajo en particular. Existen muchas y variadas formas de clasificación, cada una de las cuales enfatiza un propósito particular en la asignación de personas a categorías. Se distinguirá entre asignación, detección, certificación y selección.

La **asignación** es la distribución de personas en los diferentes programas apropiados para sus necesidades o habilidades. Por ejemplo, con frecuencia las universidades utilizan un examen de asignación en matemáticas para determinar si los estudiantes deberían inscribirse a clases de cálculo, álgebra o a un curso de regularización.

La **detección** se refiere a las pruebas o procedimientos rápidos y sencillos para identificar personas que podrían tener características o necesidades especiales. Por lo común, los psicómetras reconocen que las pruebas de detección darán por resultado muchas clasificaciones erróneas. Por tanto, se aconseja a los examinadores que realicen pruebas de seguimiento con instrumentos adicionales, antes de tomar decisiones importantes con base en pruebas de detección. Por ejemplo, para identificar a niños con un talento sumamente excepcional en pensamiento espacial, un psicólogo podría aplicar una prueba de lápiz y papel con una duración de 10 minutos, a todos los niños dentro de un sistema escolar. Después podría seleccionarse a los estudiantes cuyas puntuaciones se encontraran en el 10% superior, a fin de aplicarles una prueba más amplia.

La **certificación** y la **selección** tienen ambas una cualidad de aprobado/reprobado. Aprobar un examen de certificación confiere privilegios. Algunos ejemplos incluyen el derecho de practicar la psicología o de conducir un automóvil. Así, por lo común, la certificación

implica que una persona tiene cuando menos una destreza mínima en alguna disciplina o actividad. La selección es similar a la certificación en cuanto a que confiere privilegios, como la oportunidad de asistir a una universidad u obtener un empleo.

Otro uso de las pruebas psicológicas es el diagnóstico y la planificación de tratamiento. El **diagnóstico** consiste en dos tareas entrelazadas: determinar la naturaleza y la fuente de la conducta anormal de una persona y clasificar el patrón de conducta dentro del sistema diagnóstico aceptado. Por lo general, el diagnóstico es precursor del remedio o del tratamiento de distrés personal o ejecución deficiente.

Con frecuencia, las pruebas psicológicas desempeñan una función importante en el diagnóstico y la planificación del tratamiento. Por ejemplo, las pruebas de inteligencia son absolutamente esenciales en el diagnóstico del retraso mental. Las pruebas de personalidad son útiles para diagnosticar la naturaleza y grado de los trastornos emocionales. De hecho, algunas pruebas, como el MMPI, se diseñaron con el propósito explícito de aumentar la eficacia del diagnóstico psiquiátrico.

El diagnóstico debería ser más que una mera clasificación, más que la asignación de una etiqueta. Un diagnóstico apropiado transmite información —sobre fortalezas, debilidades, etiología y mejores opciones de regularización/tratamiento. Saber que un niño ha recibido un diagnóstico de **discapacidad para el aprendizaje** es inútil en términos generales; pero saber además que el mismo niño se encuentra en un nivel muy inferior en cuanto a comprensión de lectura, se distrae con facilidad y necesita ayuda con la fonética básica, puede proporcionar una base indispensable para la planificación del tratamiento.

Las pruebas psicológicas también pueden proporcionar una poderosa fuente de autoconocimiento. En algunos casos, la retroalimentación que recibe una persona de una prueba psicológica puede cambiar su profesión o alterar el curso de su vida. Por supuesto, no todas las situaciones de prueba psicológica proporcionan un autoconocimiento. Quizás en la mayor parte de los casos el cliente ya sabe lo que revelarán los resultados de prueba. Un estudiante universitario con un alto funcionamiento, pocas veces se sorprende de descubrir que su CI se encuentra en el rango superior. Una arquitecta no se desconcierta al escuchar que tiene excelentes habilidades de razonamiento espacial. Un estudiante con una limitada capacidad para la lectura, por lo general no se asombra de recibir un diagnóstico de "discapacidad para el aprendizaje".

Otro empleo de las pruebas psicológicas es la valoración de programas educativos y sociales. Se dirá más al respecto de la evaluación de programas educativos cuando se analicen las pruebas de aprovechamiento en un capítulo posterior. Aquí nos limitaremos en el uso de las pruebas para la evaluación de programas sociales, los cuales se diseñan para proporcionar servicios que mejoren las condiciones sociales y la vida comunitaria. Por ejemplo, el Proyecto *Head Start* es un programa con fondos federales que apoya a nivel nacional los proyectos de enseñanza preescolar para niños de sectores desprotegidos (Cicerelli, 1969; McKey y colaboradores, 1985). Lanzado en 1965 como un intento que sentara precedentes en cuanto a proporcionar programas de desarrollo infantil a familias de bajos ingresos, el *Head Start* ha suministrado enriquecimiento educativo y servicios de salud a millones de preescolares en sectores de alto riesgo.

Pero, exactamente, ¿qué impacto tiene el programa multimillonario *Head Start* sobre el desarrollo en la temprana infancia? El Congreso de EUA deseaba saber si el programa mejoraba el desempeño escolar y reducía el fracaso en la escuela de los niños inscritos. Pero los centros varían según las instituciones patrocinadoras, las características del personal, la cobertura, el contenido y los objetivos, de modo que los efectos son difíciles de evaluar. Las pruebas psicológicas proporcionan una base objetiva para responder a estas preguntas, la cual es muy superior a los informes anecdóticos o impresionistas. En general, los niños participantes en *Head Start* muestran mejorías inmediatas en CI, preparación para la escuela

y aprovechamiento académico, pero estas mejorías se disipan en los siguientes años (figura 2–2).

Hasta este punto se han analizado las aplicaciones prácticas de las pruebas psicológicas a problemas cotidianos, como la selección de personal, el diagnóstico o la evaluación de programas. En cada uno de estos casos, las pruebas satisfacen un propósito inmediato, de naturaleza práctica: ayudar al examinador a tomar decisiones sobre personas o programas. Pero las pruebas también representan una función importante en las ramas aplicada y teórica de la investigación conductual. Como ejemplo de las pruebas en la investigación aplicada, considérese el problema que enfrentan los neuropsicólogos que desean investigar la hipótesis de que la absorción de plomo a bajo nivel causa deficiencias conductuales en los niños. La única manera factible de explorar esta suposición es sometiendo a prueba a niños normales y con exposición al plomo con una batería de pruebas psicológicas. Needleman, Gunnoe, Leviton, Reed, Peresie, Maher y Barrett (1979) utilizaron un conjunto de pruebas tradicionales e innovadoras para concluir que la absorción de plomo a bajo nivel causa disminuciones en el CI, alteraciones en el tiempo de reacción y aumentos progresivos de conducta indeseable en el salón de clase. Sus conclusiones inspiraron un tumultuoso y amargo intercambio de opiniones que no se revisarán aquí (Needleman, y colaboradores, 1990). Sin embargo, las pasiones inspiradas por este estudio son la personificación de una cuestión importante: los académicos y las personas que crean la política pública respetan las pruebas psicológicas. ¿Por qué si no habrían de participar en largos y enconados debates acerca de la validez de los descubrimientos de investigación basados en pruebas?



Figura 2–2. Resultados de pruebas longitudinales del Proyecto *Head Start.*
Fuente: De McKey, R. H. y otros (1985). *The Impact of Head Start on children, families and communities.* Washington, DC:U.S. Government Printing Office. Dominio Público.

En ciertas ocasiones, las pruebas satisfacen un papel menos mundano al ayudar a los científicos a investigar cuestiones teóricas que no tienen aplicaciones prácticas inmediatas u obvias. Por ejemplo, para analizar la dependencia del campo perceptual, Witkin (1949) inventó las pruebas de la habitación inclinada y de la silla inclinada (HISI). El aparato para estas pruebas consiste en una habitación tipo caja, suspendida de pivotes con balines, de modo que pueda inclinarse en cualquier grado hacia la izquierda o derecha. Dentro de la habitación hay una silla para el sujeto, la cual también puede inclinarse con independencia de la habitación. La tarea del sujeto es llevar su cuerpo a una posición que se perciba como recta. Los sujetos que dependen del campo alinean de cierto modo sus cuerpos con respecto a la habitación, en vez de basarse en la fuerza de gravedad percibida. Los sujetos independientes del campo se ven menos afectados por la habitación mal alineada y están más a tono con sus señales perceptuales internas; es decir, sus juicios perceptuales son relativamente independientes de la información visual distorsionada. La HISI inspiró toda una vida de investigaciones acerca del desarrollo de la personalidad, pero en raras ocasiones se aplicó a algún problema práctico de prueba.

QUIÉN PUEDE OBTENER LAS PRUEBAS

Por lo general, aquellos que desarrollan las pruebas, los editores y los examinadores en psico-

logía proporcionan las pruebas psicológicas sólo a personas calificadas que tengan una necesidad legítima de estudiar o utilizar estos materiales. Existen tres razones por las que se restringe el acceso a las pruebas psicológicas:

1. En manos de personas no calificadas, pueden causar daño.
2. El proceso de selección resulta inválido para personas que pueden adelantarse a las preguntas de la prueba.
3. La filtración del contenido del reactivo al público en general destruye la eficacia de la prueba.

Se analizarán con más detalle cada uno de estos puntos.

Un examinador no calificado puede cometer errores en la selección, aplicación, calificación o interpretación de las pruebas psicológicas, lo cual puede dañar al sujeto. Las posibilidades de error y daño son siempre ilimitadas, de modo que aquí sólo se proporcionará un ejemplo. Un error común entre los examinadores sin experiencia consiste en no dar crédito a una persona de mayor edad por los reactivos más fáciles, no aplicados, de una subescala. Por ejemplo, en una subescala hipotética de inteligencia con 20 reactivos, el manual de la prueba podría especificar que a un individuo mayor sólo se le deberían aplicar los reactivos del 11 al 20, con la suposición de que los reactivos más fáciles (1 a 10) seguramente se contestarán de manera correcta. Sin embargo, las instrucciones podrían especificar que el examinado debería recibir acreditación en puntos por los reactivos 1 a 10. No agregar estos puntos produciría que la puntuación fuera drásticamente baja, con consecuencias negativas para la persona examinada.

Otra razón para limitar la disponibilidad de las pruebas es que el acceso ilícito a su contenido socava la efectividad del proceso de selección. Dicho en términos sencillos, los individuos examinados que han tenido acceso previo a una prueba de selección pueden aprender a producir los resultados de prueba deseados. Por último, debería ser obvio que la filtración de los reactivos de prueba al público produce que la prueba sea totalmente inútil. Si los individuos pueden memorizar las respuestas a las preguntas de la prueba, su desempeño mejorará de manera artificial. Para dar un ejemplo extremo, una persona invidente podría aprobar una prueba de visión del color a través de memorizar las respuestas correctas.

FUENTES DE INFORMACIÓN SOBRE PRUEBAS

Un libro de texto sobre pruebas psicológicas no puede analizar todos los instrumentos que tengan un interés potencial para los lectores. ¡Simplemente hay demasiadas pruebas! Lo que es más, cada año se desarrollan docenas de pruebas nuevas y útiles. El estudiante serio de las pruebas psicológicas necesitará pautas y estrategias para aprender sobre dichos métodos, no una lista estática de recomendaciones.

La información sobre pruebas psicológicas puede obtenerse de cuatro fuentes: libros de referencia, catálogos editoriales, revistas y manuales de prueba. Se citarán unos cuantos ejemplos destacados de cada categoría, pero se espera que el lector asimile una estrategia para adquirir conocimientos, en vez de depender en forma exclusiva de estas citas específicas.

La mejor fuente de referencia en cuanto a información sobre las pruebas establecidas es el *Mental Measurements Yearbook* (MMY), que publica el *Buros Institute for Mental Measurement*, en la *University of Nebraska*. Periódicamente se publica una nueva edición del MMY (Buros, 1978; Conoley e Impara, 1995; Conoley y Kramer, 1989, 1992; Mitchell, 1985). El MMY incluye reseñas críticas de las pruebas y una lista de referencias importantes.

La *Test Corporation of America* publica varios libros de referencia y críticas excelentes sobre pruebas. *Tests*, editado por Sweetland y Keyser (1987), es una lista amplia de instrumentos de evaluación en psicología, educación y empresas. *Test Critiques* (volúmenes I–VI), editado por Keyser y Sweerland (1984–1988), proporciona evaluaciones a profundidad de las

pruebas psicológicas, educativas e industriales. Además, el mismo editor ofrece una serie de libros sobre pruebas con niños, adolescentes, adultos y personas de edad avanzada (Weaver, 1984; Harrington, 1986; Swiercinsky, 1985). Otra excelente fuente de medidas clínicas cortas y sencillas es *Measures for Clinical Practice: A Sourcebook*, de Corcoran y Fisher (1994). En este compendio, se proporcionan referencias cruzadas de instrumentos para adultos, niños, parejas y familias, según sea el área problemática.

Otra manera de conocer las pruebas es pedir los catálogos de los principales editores de pruebas. El Apéndice B (Direcciones de Editoriales de Pruebas) lista los nombres y direcciones de las editoriales y distribuidores de pruebas más destacados en Estados Unidos. El Apéndice C (Principales Pruebas y sus Editores) proporciona una lista categorizada de pruebas notables y de las editoriales que las publican. El último MMY contiene un directorio más amplio de editoriales y pruebas.

Muchas revistas psicológicas publican artículos sobre la confiabilidad y validez de las pruebas más conocidas. La mejor manera de localizar los estudios sobre una prueba específica es a través de los *Psychological Abstracts*, un destacado compendio mensual de las principales revistas en psicología y campos relacionados. En general, las principales bibliotecas tienen un PsychLit en CD-ROM. Ésta es la versión para computadora de los *Psychological Abstracts*. Si el lector sólo está interesado en examinar los estudios recientes sobre pruebas psicológicas, sería adecuado consultar los últimos números de las siguientes revistas:

Assessment
Advances in Personality Assessment
The Journal of Psychoeducational Assessment
Psychology in Schools
The Journal of School Psychology
Educational and Psychological Measurement
Psychological Assessment
The Journal of Clinical Neuropsychology
The Journal of Clinical Psychology
The Journal of Personality Assessment

Por último, una importante e ignorada fuente de información sobre cualquier prueba específica es su manual. Un buen manual contiene información esencial sobre normas, estandarización, aplicación, confiabilidad y validez.

RESUMEN

1. Una prueba puede definirse como un procedimiento estandarizado para tomar una muestra de conducta y describirla con categorías y puntuaciones. Además, la mayoría de las pruebas tienen normas o estándares con base en los cuales pueden utilizarse los resultados para pronosticar otras conductas más importantes.
2. Las pruebas siempre constituyen una muestra de la conducta, nunca la totalidad de aquello que el examinador busca medir. Por tal razón, los resultados de prueba siempre incorporan cierto grado de error de medición.
3. En una prueba con referencia a la norma, la puntuación de prueba del individuo se interpreta en relación con las puntuaciones obtenidas por otras personas en la misma prueba. En una prueba con referencia al criterio, el énfasis se coloca sobre aquello que la persona examinada puede hacer con respecto a criterios educativos definidos de manera estrecha.
4. La evaluación es el proceso de recopilar información sobre una persona y utilizarla para hacer inferencias acerca de sus características o para predecir su conducta. La evaluación incorpora las pruebas, pero

es más amplia y puede incluir observaciones, entrevistas y otras fuentes de información.

5. Las pruebas grupales son medidas de lápiz y papel adecuadas para examinar a grandes grupos de personas a un mismo tiempo. Las pruebas individuales están diseñadas para una aplicación a una sola persona; con ello, el examinador puede observar la motivación y otras características de la persona examinada.
6. Una clasificación arbitraria, pero útil, de las pruebas psicológicas es la siguiente: de inteligencia, aptitud, aprovechamiento, creatividad, personalidad, intereses, conductual y neuropsicológica. Las características de estas pruebas se resumen en el cuadro 2–1.
7. Se pueden distinguir cinco usos de las pruebas: clasificación, diagnóstico y planificación del tratamiento, autoconocimiento, valoración de programas e investigación.
8. La clasificación puede descomponerse aún más en: asignación, que es la clasificación de personas a programas apropiados; detección, que es la identificación rápida de personas con características o necesidades especiales; certificación (p.ej. para obtener una licencia de manejo) y selección (p. ej., para una universidad).
9. El acceso a las pruebas psicológicas está controlado de manera estricta, de modo que sólo las personas con el entrenamiento apropiado puedan tener acceso a ellas. Muchos editores de pruebas las dividen en tres niveles de complejidad, las cuales requieren grados de pericia cada vez mayores para su aplicación.
10. Las fuentes de información sobre pruebas incluyen la serie del *Mental Measurements Yearbook* y los volúmenes de las *Test Critiques*. Algunas revistas como *Assessment* y *The Journal of Psychoeducational Assessment* también proporcionan información sobre pruebas psicológicas.

TÉRMINOS Y CONCEPTOS CLAVE

psicómetra	p. 36	pruebas de creatividad	p. 42
prueba	p. 36	creatividad	p. 42
procedimiento estandarizado	p. 37	pruebas de personalidad	p. 43
normas	p. 39	inventarios de intereses	p. 43
muestra de estandarización	p. 39	procedimientos conductuales	p. 43
prueba referida a la norma	p. 39	pruebas neuropsicológicas	p. 43
prueba referida a criterio	p. 39	clasificación	p. 45
evaluación	p. 40	asignación	p. 45
pruebas grupales	p. 40	detección	p. 45
pruebas individuales	p. 40-41	certificación	p. 45
pruebas de inteligencia	p. 41	diagnóstico	p. 46
prueba de aptitud	p. 42	discapacidad para el aprendizaje	p. 46
pruebas de aprovechamiento	p. 42		

TEMA 2B
Proceso de aplicación

Ejemplo de caso 2–1. Impacto de las pruebas no estandarizadas
Procedimientos estandarizados en la aplicación de pruebas
Procedimientos deseables en la aplicación de pruebas
Influencia del examinador
Antecedentes y motivación de la persona examinada
Aspectos sobre calificación
Resumen
Términos y conceptos clave

Las pruebas psicológicas son un proceso dinámico bajo la influencia de muchos factores. Aunque los examinadores se esfuerzan en garantizar que los resultados de prueba reflejen de manera precisa los rasgos o capacidades evaluados, pueden influir muchos factores externos en el resultado del examen psicológico. En esta sección se revisará el impacto potencialmente crucial de varias fuentes de influencia: el modo de aplicación, las características del examinador, el contexto de la prueba, la motivación y experiencia de la persona examinada y el método de calificación.

La sensibilidad del proceso de prueba a las influencias externas es obvio en esos espectáculos poco comunes y atroces de las pruebas no estandarizadas que se presentan de vez en cuando (ejemplo de caso 2–1). Sin embargo, los resultados inválidos de prueba no se originan sólo de fuentes obvias, como la aplicación evidentemente no estandarizada, el examinador hostil, la habitación ruidosa de prueba, el individuo asustado o la calificación descuidada. Además, existen numerosas y sutiles maneras en las que el método, el examinador, el contexto, la motivación o la calificación pueden alterar los resultados de prueba. Se proporciona un examen detallado de estas influencias externas en lo que resta de este tema.

PROCEDIMIENTOS ESTANDARIZADOS EN LA APLICACIÓN DE PRUEBAS

La interpretación de una prueba psicológica es más confiable cuando las mediciones se obtienen bajo las condiciones estandarizadas descritas en el manual de prueba que publica su editor. Los procedimientos no estandarizados pueden alterar el significado de los resultados de la prueba, que pueden resultar inválidos y, por tanto, engañosos. Los procedimientos estandarizados son tan importantes que se enumeran como un criterio esencial para el examen válido en los *Standards for Educational and Psychological Testing* (1985, 1999), un manual de referencia que publican de manera conjunta la *American Psychological Association* y otros grupos:

> En las aplicaciones típicas, quienes aplican una prueba deben seguir de manera cuidadosa los procedimientos estandarizados para la aplicación y calificación que especifica el editor de la prueba. Las especificaciones concernientes a las instrucciones para las personas examinadas, límites de tiempo, forma de presentación o respuesta del reactivo y materiales o equipo de prueba deben observarse de manera estricta. Sólo deben hacerse excepciones con base en el juicio profesional considerado de modo cuidadoso, principalmente en las aplicaciones clínicas (AERA, APA, NCME, 1985)

Supóngase que las instrucciones para la sección de vocabulario de una prueba de inteligencia para niños especifica que el examinador debe preguntar: "¿qué significa la palabra **diván**, qué es un diván?" Si el sujeto respondiera "nunca he escuchado esa palabra", un examinador sin experiencia podría sentirse tentado a responder: "ya sabes, un sofá —¿qué es un sofá?" Esto podría parecerle al lector como una forma inocente de juego limpio, un simple refraseo de la pregunta original. Sin embargo, al alejarse de los procedimientos estandarizados, el examinador ha aplicado en realidad otra prueba. El punto de pedir la definición de **diván** (y no **sofá**) es precisamente que **diván** es más difícil de definir y, por tanto, es un mejor índice de habilidades de nivel superior en cuanto a vocabulario.

Aunque los procedimientos estandarizados de prueba son normalmente esenciales, existen casos en que es deseable, o incluso necesaria, la flexibilidad en los procedimientos. Como lo sugieren las normas de la APA, dichas desviaciones deben ser razonadas y deliberadas. Aquí es pertinente realizar una analogía entre el espíritu y la letra de la ley. Un examinador demasiado entusiasta podría captar, por un decir, la letra de la ley al adherirse de manera literal y estricta a los procedimientos de prueba expresados en el manual de edición. Pero, ¿en realidad ésta era la intención del editor? ¿Siquiera es la manera en que, en realidad, se aplicó la prueba a la muestra normativa? Es más probable que los editores preferirían que los examinadores captaran el espíritu de la ley incluso si, en ocasiones, fuera necesario adaptar un tanto los procedimientos de la prueba.

Considérese la siguiente situación que surgió cuando un psicólogo le aplicó una prueba estandarizada de inteligencia a una estudiante universitaria ansiosa y demasiado concreta. Cuando se le preguntó: "¿cuántos son cuatro dólares más cinco dólares?", la estudiante respondió: "cuatro dólares son cuatro dólares y cinco dólares son cinco dólares". Una interpretación literal del manual de prueba requeriría que el examinador registrara un crédito de cero y prosiguiera con el siguiente reactivo. Sin embargo, la pregunta tenía el propósito de examinar las habilidades aritméticas y no lo concreto del pensamiento. Así, el examinador hizo de nuevo la pregunta con un ligero cambio en cuanto a énfasis: "¿cuántos son cuatro dólares y cinco dólares?" La sujeto lanzó una fuerte carcajada y respondió de inmediato "nueve dólares —no me percaté de que era una pregunta de aritmética".

La necesidad de adaptar los procedimientos estandarizados para una prueba es particularmente evidente cuando se examina a personas con ciertos tipos de discapacidades. Se podría permitir que un individuo con una discapacidad del habla escribiera sus respuestas a preguntas presentadas de manera oral o que utilizara

EJEMPLO DE CASO 2-1
IMPACTO DE LAS PRUEBAS NO ESTANDARIZADAS

Un estudiante de posgrado en psicología, con el cabello descuidado y crespo y barba completa, se acerca a una niña de seis años que se encuentra en la sala de espera y le solicita que lo siga. El estudiante de posgrado habla de manera brusca con un acento de Brooklin que es desconocido para la pequeña. Ésta se muestra visiblemente atemorizada y permanece sentada hasta que su madre la empuja para que se levante, mientras le lanza la no muy tranquilizadora advertencia: "Apúrate o tu papi se va a enojar mucho". Son la 8:30 de la noche y la niña acaba de terminar con una batería completa de pruebas médicas. Tan sólo una hora antes, una enfermera introdujo una enorme aguja en el brazo derecho de la niña y le extrajo una cantidad de sangre equivalente a una jeringa. De manera obediente, pero con titubeos, la niña sigue al examinador a una habitación apenas iluminada. El constante ruido del tráfico en una vía rápida cercana le hace difícil escuchar las instrucciones del examinador. Éste se encuentra cansado —como parte de un proyecto de investigación epidemiológica de grandes proporciones, ha estado examinando a niños desde las 7:30 de la mañana. Sin embargo, le aplica la Escala Wechsler de Inteligencia para Niños (WISC) a un paso acalorante, terminando la prueba en 25 minutos. Después, esa misma noche califica la prueba en su habitación de un motel. Por descuido, no le da crédito a la niña por varios reactivos simples y también comete un error administrativo al sumar las puntuaciones estándar. Se informa que la niña tiene un CI de 69 puntos, lo cual posiblemente indica retraso mental leve. Por fortuna, al día siguiente el supervisor de pruebas detecta los errores de calificación y cuestiona al estudiante de posgrado sobre las condiciones de la prueba. Al reconocer que los resultados no son válidos, el supervisor vuelve a aplicar la prueba un mes después bajo condiciones favorables; la pequeña obtiene puntuaciones mucho más altas, dentro del rango promedio bajo. Al comparar ambos protocolos de prueba, el examinador de posgrado aprende una valiosa lección: las pruebas válidas son más que una simple cuestión de hacer preguntas, registrar respuestas y sumar las puntuaciones de un modo despreocupado.

ademanes o pantomima como respuesta para algunos reactivos. Por ejemplo, una pregunta de prueba podría ser: "¿qué forma tiene una pelota?" La pregunta está diseñada para sondear el conocimiento de la persona acerca de formas comunes, no para examinar si ésta puede verbalizar la palabra "redonda". Escribir como respuesta la palabra **redonda** y dar una respuesta con ademanes (un movimiento circular con el dedo índice) son, también, igualmente correctas.

Los ajustes menores a los procedimientos, que se adhieren al espíritu en el que se desarrolló la prueba, ocurren de manera regular y no son causa de alarma. Estas adaptaciones menores no invalidan las normas establecidas —por el contrario, la adaptación apropiada de los procedimientos es necesaria de modo que las normas puedan continuar siendo válidas. Después de todo, los examinadores que recolectaron datos sobre la muestra de estandarización no actuaron como autómatas sin corazón cuando presentaron las preguntas a los sujetos. Los examinadores que deseen obtener resultados válidos deben ejercer, del mismo modo, una flexibilidad razonada en los procedimientos de prueba.

Sin embargo, es necesario tener una considerable experiencia clínica para determinar si un ajuste en el procedimiento es menor o tan sustancial que las normas existentes ya no pueden aplicarse. Ésta es la razón por la cual los examinadores en psicología pasan, por lo común, por una extensa experiencia supervisada antes

de que se les permita aplicar e interpretar pruebas individuales de capacidad o de personalidad.

En ciertos casos, un examinador se alejará a un grado considerable de los procedimientos estandarizados a sabiendas de lo que hace; esta práctica impide el uso de las normas disponibles para la prueba. En estos casos, la prueba se utiliza para ayudar a formular juicios clínicos, más que para determinar un índice cuantitativo. Por ejemplo, cuando se examina a pacientes con afasia, podría ser deseable ignorar por completo los límites de tiempo y aceptar las respuestas indirectas (Eisenson, 1954). Incluso el examinador podría no calcular una puntuación. En estos casos poco comunes la prueba se vuelve, en efecto, un complemento de la entrevista clínica. Por supuesto, cuando el examinador no se adhiere a los procedimientos estandarizados, ello debería especificarse de manera explícita en el informe por escrito.

PROCEDIMIENTOS DESEABLES EN LA APLICACIÓN DE PRUEBAS

Se podría escribir un pequeño tratado acerca de los procedimientos deseables en la aplicación de pruebas, pero tendremos que conformarnos con un breve listado de los puntos más esenciales. Para mayores detalles, el lector interesado puede consultar a Sattler (1988) sobre pruebas individuales con niños y a Clemans (1971) sobre pruebas grupales. Se analizarán primero las pruebas individuales y después se enumerarán de manera breve algunos puntos importantes acerca de los procedimientos deseables en las pruebas con grupos.

Un componente esencial de las pruebas individuales es que los examinadores deben familiarizarse íntimamente con los materiales e instrucciones antes de comenzar con la aplicación. En su mayoría, esto implica extensos ensayos y anticipación de las circunstancias poco comunes y de la respuesta apropiada. Un examinador bien preparado tiene memorizados los elementos clave de las instrucciones verbales y está listo para manejar lo inesperado.

Con frecuencia, el estudiante de evaluación inexperto supone que los procedimientos de examen son tan sencillos y simples que una sola lectura rápida del manual será suficiente como preparación para la prueba. Aunque algunas pruebas individuales son sumamente rudimentarias y poco complicadas, muchas de ellas tienen aspectos complejos de aplicación que, si no se toman en cuenta, pueden causar que el individuo examinado falle de manera innecesaria en los reactivos. Por ejemplo, Choi y Proctor (1994) encontraron que 25 entre 27 estudiantes de posgrado cometían serios errores en la aplicación de la Stanford-Binet: Cuarta Edición, incluso aunque se videogrababan las sesiones y los estudiantes sabían que se evaluarían sus habilidades de prueba. La atención apropiada a los detalles de aplicación es esencial para obtener resultados válidos.

La necesidad de tener una íntima familiaridad con los procedimientos de prueba se ilustra muy bien en la subprueba de Diseño con Cubos del WAIS-III (Wechsler, 1997). Los materiales para la subprueba incluyen nueve bloques (cubos) de color rojo en dos lados, blanco en dos lados y rojo y blanco en dos lados. La tarea de la persona evaluada consiste en utilizar los cubos para construir patrones representados en tarjetas. Para los diseños iniciales se necesitan cuatro cubos, mientras que en los diseños más difíciles se utilizan los nueve cubos (figura 2–3).

Los individuos inteligentes no tienen dificultad para comprender esta tarea y las instrucciones exactas no influyen de manera apreciable en su desempeño. Sin embargo, las personas cuya inteligencia es promedio o inferior a éste requieren de las elaboradas demostraciones y correcciones que se especifican en el Manual del WAIS-III (Wechsler, 1997). En particular, el examinador muestra los primeros dos diseños y responde al éxito o fracaso del examinado en ellos, según un flujo complejo de reacción y contrarreacción como se presenta en las **tres páginas** de instrucciones. Lo sentimos por el examinador que no ha ensayado esta subprueba y anticipado la respuesta apropiada para las personas que fallan en los primeros dos diseños.

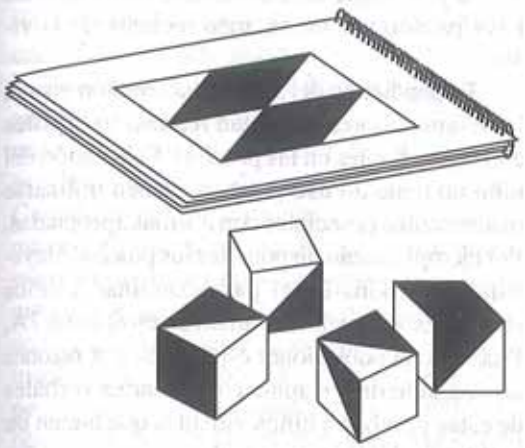


Figura 2–3. Materiales similares a la subprueba de Diseños con Cubos del WAIS-III.

SENSIBILIDAD HACIA LAS DISCAPACIDADES

Otro ingrediente importante de la aplicación válida de una prueba es la sensibilidad a las discapacidades de la persona evaluada. Las discapacidades en audición, visión, habla o control motor pueden distorsionar gravemente los resultados de prueba. Si el examinador no reconoce la discapacidad física responsable del desempeño deficiente en la prueba, puede clasificarse al sujeto como discapacitado a nivel intelectual o emocional cuando, de hecho, el problema esencial es una discapacidad sensorial o motora.

Vernon y Brown (1964) informaron sobre el trágico caso de una niña pequeña a la que se le relegó a un hospital para individuos con retraso mental, como consecuencia de la insensibilidad del examinador hacia una discapacidad física. El examinador no se dio cuenta de que la niña era sorda y concluyó que el CI de 29 en la Stanford-Binet era válido. La niña permaneció en el hospital durante cinco años, pero se le dio de alta después de que obtuvo ¡un CI de 113 en una prueba de inteligencia basada en el desempeño! Después de dejar el hospital, ingresó a una escuela para sordos y logró buen progreso.

Las personas con discapacidades pueden requerir pruebas especializadas para la evaluación válida. El lector encontrará una extensa discusión de las pruebas disponibles para individuos excepcionales en el tema 7A, Pruebas con poblaciones especiales. En esta sección nos concentraremos en los preocupantes problemas que surgen cuando se utilizan pruebas estandarizadas para poblaciones normales con sujetos que tienen discapacidades leves o moderadas. Incluimos análisis por separado del proceso de prueba con individuos que tienen un problema de audición, visión, habla o control motor. Sin embargo, el lector necesita saber que muchas personas excepcionales tienen múltiples discapacidades (Tweedie y Shroyer, 1982).

Las pruebas válidas con sujetos que tienen discapacidades auditivas requieren antes que nada que el examinador ¡detecte la existencia de la discapacidad! Con frecuencia esto es más difícil de lo que parece. Muchas personas con una pérdida auditiva leve aprenden a compensar su discapacidad fingiendo que comprenden lo que otros dicen y esperando que las siguientes señales dentro de la conversación les ayuden a aclarar las palabras o frases percibidas apenas. Como resultado, otras personas —incluyendo los psicólogos— pueden no darse cuenta de que el individuo con una alteración auditiva leve tiene alguna discapacidad.

La incapacidad para percatarse de una pérdida auditiva constituye un problema en particular para los individuos jóvenes, quienes, por lo general, son malos informantes de su discapacidad. Los niños pequeños también están propensos a pérdidas auditivas fluctuantes, debido a la acumulación periódica de líquido en el oído medio durante intervalos de enfermedad leve (Vernon y Alles, 1986). Un niño con una pérdida auditiva fluctuante puede tener audición normal en la mañana, pero sólo unas cuantas horas después percibe el habla dentro de una conversación como si fuera un susurro.

Las indicaciones de una posible dificultad auditiva incluyen falta de respuesta normal ante el sonido, falta de atención, dificultades para

obedecer instrucciones orales, observación intensa de los labios del hablante y articulación deficiente (Sattler, 1988). En todos los casos donde se sospecha la existencia de una alteración auditiva, es crucial canalizar al sujeto a un examen audiológico. Si se confirma un problema auditivo grave, entonces el examinador debería considerar el uso de una de las pruebas especializadas que se analizan en el tema 7A, Pruebas con poblaciones especiales. En el caso de personas con una pérdida auditiva leve, es esencial que el examinador se coloque de frente al sujeto, hable más fuerte y repita las instrucciones lentamente. También es importante localizar una habitación silenciosa para la prueba. De manera ideal, dicha habitación tendrá cortinas y superficies con textura en las paredes para reducir al mínimo los efectos distractores de los ruidos de fondo.

En contraste con los individuos que tienen alteraciones auditivas, las personas con discapacidades visuales generalmente atienden bien a los materiales de prueba presentados de modo verbal. La persona con alteraciones visuales introduce un tipo diferente de reto para el examinador; detectar la existencia de una alteración visual y después, asegurarse de que el sujeto puede ver bien los materiales de prueba.

Detectar las alteraciones visuales es una cuestión sencilla en los sujetos adultos —en la mayoría de los casos, un individuo maduro informará libremente que tiene una alteración visual, en especial si se le pregunta. Sin embargo, los niños son malos informantes de sus capacidades visuales, de modo que los examinadores necesitan conocer los signos y síntomas de una posible alteración visual en niños pequeños. El sentido común es un buen punto de inicio: los niños que entrecierran los ojos, parpadean de manera excesiva o pierden el lugar en la lectura, pueden tener un problema de visión. Sostener los libros o los materiales de prueba cerca de los ojos es otro signo de sospecha. La visión borrosa o doble puede significar problemas visuales, como los dolores de cabeza o las náuseas después de leer. En general, es tan común que los niños requieran anteojos correctores, que los examinadores deben estar alertas a un problema de visión en cualquier sujeto joven que no utilice anteojos y que no haya pasado por un examen reciente de la visión.

Dependiendo del grado de alteración visual, los examinadores necesitan realizar los ajustes correspondientes en las pruebas. Si la visión del niño no tiene un uso práctico, deben utilizarse instrumentos especiales con normas apropiadas. Por ejemplo, están disponibles las pruebas Hays-Binet y Perkins-Binet para examinar a niños ciegos. Estas pruebas se analizan en el tema 7A, Pruebas con poblaciones especiales. Por razones obvias, sólo deben aplicarse las partes verbales de estas pruebas a niños videntes que tienen un problema visual no corregido.

Las alteraciones del habla representan otro problema para quienes deben hacer un diagnóstico. Las respuestas verbales de las personas con alteraciones del habla son difíciles de descifrar. Debido a la incapacidad del examinador para entender las respuestas, los sujetos pueden recibir un menor crédito del que merecen. Sattler (1988) relata el lamentable caso de Daniel Hoffman, un jovencito con alteraciones del habla que pasó toda su niñez en clases para individuos con retraso mental debido a que su CI en la Stanford-Binet fue de 74 puntos. En realidad, su inteligencia se encontraba dentro del rango normal, como lo revelaron otras pruebas basadas en el desempeño. En otro error trágico de evaluación, a un paciente en Inglaterra se le confinó equivocadamente a un pabellón para personas con retraso mental grave, debido a que su parálisis cerebral producía que su discurso fuera incomprensible. El paciente se encontraba limitado a una silla de ruedas y casi no tenía control motor, por lo que su desempeño en las pruebas no verbales también estaba sumamente deteriorado. El personal supuso que tenía un retraso grave, así que el paciente permaneció durante décadas en el pabellón más remoto. Sin embargo, se hizo amigo de un compañero residente que podía comprender la interpretación gutural que el paciente hacía del abecedario. El amigo tenía un retraso grave, sin embargo, podía reconocer las teclas en una máquina de escribir. Con un laborioso esfuerzo para escribir

letra por letra, el paciente con parálisis cerebral incapacitante escribió y publicó su autobiografía, utilizando a su amigo con retraso mental como un conducto hacia el mundo real.

Incluso si su discapacidad es leve, las personas con parálisis cerebral u otras alteraciones motoras pueden recibir penalización en pruebas de ejecución con límite de tiempo. Cuando se somete a prueba a una persona con una discapacidad motora leve, los examinadores podrían omitir las subpruebas de ejecución cronometradas o descontar estos resultados si son consistentemente más bajos que las puntuaciones de las subpruebas sin límite de tiempo Si un sujeto tiene una discapacidad motora obvia —como lo es una dificultad para manipular las piezas de un rompecabezas— entonces los instrumentos estándar aplicados de manera normal son generalmente inapropiados. Se han desarrollado de manera expresa varios instrumentos alternativos para examinar a personas con parálisis cerebral y otras alteraciones motoras y las pruebas estándar se han adaptado de modo ingenioso y se han establecido nuevas normas para estas adaptaciones (tema 7A, Pruebas con poblaciones especiales).

PROCEDIMIENTOS DESEABLES CON PRUEBAS DE GRUPO

Por lo común, los psicólogos y educadores suponen que casi cualquier adulto puede aplicar de manera adecuada las pruebas grupales, en tanto tenga el manual requerido. La aplicación de una prueba grupal parecería un procedimiento simple y sencillo en el que se pasan los formatos y lápices, se leen las instrucciones, se toma el tiempo y se recogen los materiales.

En realidad, conducir una prueba grupal requiere tanto refinamiento como aplicar una prueba individual, un punto que reconoció hace años Traxler (1951):

> Es dudoso que cualquier institución educativa confiaría la aplicación y calificación de la Escala Stanford-Binet a cualquier persona que no fuera un psicómetra entrenado, pero existe la impresión bastante general entre quienes manejan las escuelas de que casi cualquiera puede aplicar y calificar una prueba grupal, con sólo tener un manual junto de sí. Sin embargo, de hecho, la aplicación de una prueba grupal a una clase de alumnos es, en cierto sentido, un procedimiento más riguroso que aplicar una prueba individual a un solo alumno. La rutina es más rígida y la penalidad para los errores se multiplica por el número de individuos dentro del grupo. En la aplicación de una prueba individual se puede permitir un cierto grado de libertad de acción, a fin de crear una situación favorable para obtener las respuestas de un individuo en particular; pero cuando se somete a prueba un grupo, es imperativa la completa fidelidad hacia todos los detalles del procedimiento prescrito, si es que el examinador quiere evitar la pérdida de tiempo de muchos individuos y si los resultados deben tener el mismo significado para todos.

Existen muchas maneras en que la aplicación y calificación negligentes pueden perjudicar los resultados de una prueba grupal, causando un sesgo para el grupo completo o afectando sólo a ciertos individuos. A continuación se resumirán sólo las deficiencias y los errores más importantes y referimos al lector a Traxler (1951) y Clemans (1971) para un análisis más completo.

Sin duda, la más grande fuente de error en la aplicación de una prueba grupal tiene que ver con tomar de manera inadecuada el tiempo en las pruebas cronometradas. Los examinadores deben conceder el tiempo suficiente para el proceso completo de prueba: organización, lectura de las instrucciones en voz alta y presentación de la prueba en sí por parte de los sujetos. Conceder el tiempo suficiente requiere previsión en la programación. Por ejemplo, en muchos ambientes escolares, los niños deben proceder a la siguiente clase en un horario establecido, sin considerar las actividades en proceso. Los examinadores sin experiencia podrían sentirse tentados a reducir el límite de tiempo designado para una prueba, de modo que se pueda respetar el horario de la escuela. Por supuesto, reducir el tiempo en una prueba causa que las normas sean completamente inválidas y con toda probabilidad reduce la puntuación de la mayoría de los sujetos en el grupo.

Conceder demasiado tiempo para una prueba puede constituir también un tremendo error.

Por ejemplo, considérese el impacto de recibir tiempo adicional en la *Miller Analogies Test* (MAT; Prueba de Analogías de Miller), una prueba de razonamiento de alto nivel que en algún tiempo requirieron muchas universidades para la solicitud de ingreso al posgrado. Dado que la MAT es una prueba con límite de tiempo que necesita de pensamiento analógico rápido, conceder más tiempo permitiría que la mayoría de los individuos examinados resolvieran varios problemas adicionales. Es probable que este tipo de error de prueba disminuiría la validez de los resultados de la MAT, como mecanismo de predicción del desempeño en el posgrado.

Una segunda fuente de error en la aplicación de pruebas grupales es la falta de claridad en las instrucciones para los examinados. Los examinadores deben leer las instrucciones con lentitud, con una voz clara y fuerte que atraiga la atención de los sujetos. Las instrucciones no deben parafrasearse. Cuando el manual lo permite, los examinadores deben detenerse en la lectura y aclarar las dudas de los individuos que tienen alguna confusión.

Las variaciones en las condiciones físicas de aplicación de prueba constituyen una tercera fuente de error potencial en la conducción de una prueba grupal. Los examinadores deben asegurarse de que la habitación de examen esté bien iluminada y, de ser necesario, que tenga calefacción o aire acondicionado para controlar las variaciones extremas en cuanto a temperatura y humedad. Clemans (1971) ha señalado que es poco común que los autores de pruebas entren en detalles sobre las especificaciones relativas a iluminación, temperatura y humedad, dado que los examinadores y sujetos, con unas cuantas excepciones, tendrán que tolerar las condiciones existentes. Sin embargo, es obvio que los examinandos no pueden tener un desempeño óptimo si se les somete a prueba en una habitación con iluminación pobre o que es demasiado fría u opresivamente caliente y húmeda. Los examinadores previsores deberían hacerle el favor a sus examinados de programar que las pruebas grupales importantes se presenten en un ambiente agradable y bien iluminado.

La calidad de la superficie de escritura puede tener una importancia crucial para las pruebas grupales válidas, en especial con niños. La afirmación de Traxler (1951) de que las escuelas varían ampliamente en sus instalaciones para la aplicación de pruebas grupales, todavía es válida hoy en día:

> Tan sólo en la cuestión del espacio para escribir, algunas escuelas emplean mesas grandes y cómodas, otras usan pupitres, otras sillas con brazos e incluso otras aplican sus pruebas en un auditorio, en el que cada alumno escribe en un "escritorio" portátil de cartón o aun sobre sus rodillas. No es razonable esperar resultados totalmente comparables bajo tales condiciones variables.

La importancia de la superficie de escritura se aumenta por la tendencia actual a utilizar hojas separadas de respuesta. Los sujetos necesitan de un espacio más amplio de escritorio cuando emplean hojas separadas de respuesta que bajo otras condiciones. Aunque pocos editores de prueba lo señalan, sería bueno especificar en los manuales de prueba las variaciones admisibles en cuanto a superficie de escritura que de todos modos permiten resultados comparables de prueba.

El ruido es otro factor que debe controlarse en las pruebas grupales. Durante algún tiempo se ha sabido que el ruido provoca una disminución en el desempeño, en especial para tareas de alta complejidad (p. ej. Boggs y Simon, 1968). Es de sorprender que exista poca investigación acerca de los efectos del ruido en las pruebas psicológicas. Sin embargo, parece casi seguro que el ruido fuerte, en especial si es intermitente e imposible de predecir, causará que las puntuaciones de prueba se reduzcan de manera sustancial. No se puede esperar que los niños en una escuela primaria tengan un buen desempeño mientras que un obrero de la construcción golpea con un marro una pared de cemento en la habitación contigua. Para ser justos con los examinandos, existen ocasiones en las que debe reprogramarse la aplicación de la prueba.

Una cuarta fuente de error en la aplicación de una prueba grupal es la incapacidad para

explicar cuándo pueden adivinarse las respuestas y si está permitido hacerlo. Quizá con más frecuencia que cualquier otra duda, a los examinadores se les pregunta si "¿existe alguna sanción por adivinar mal una respuesta?" En la mayoría de los casos, quienes desarrollan pruebas anticipan esta cuestión y proporcionan pautas explícitas para los sujetos en cuanto a las ventajas, desventajas, o ambas, de conjeturar una respuesta. Los examinadores no deben dar consejo adicional sobre adivinar las respuestas —ello constituiría una grave desviación con respecto al procedimiento estandarizado.

La mayoría de quienes desarrollan pruebas incorporan una **corrección de conjeturas** basada en principios establecidos de probabilidad. Considérese una prueba de opción múltiple que tiene cuatro alternativas por reactivo. En aquéllos donde el sujeto realiza una conjetura aleatoria, sin tener el conocimiento, las probabilidades de estar en lo correcto son de 1 entre 4, mientras que las probabilidades de estar mal son de 3 entre 4. Así, por cada tres conjeturas incorrectas, habrá una correcta que refleje la suerte más que el conocimiento. Supóngase que una niña responde correctamente a 35 preguntas de una prueba de 50 reactivos, pero en nueve preguntas sus respuestas son erróneas. En total ha respondido a 44 preguntas, dejando seis sin responder. El hecho de que haya elegido la alternativa incorrecta en nueve preguntas sugiere que ha obtenido tres respuestas correctas por suerte más que por conocimiento. Recuérdese que, en el caso de las conjeturas aleatorias, se espera que haya, en promedio, tres respuestas erróneas por cada respuesta correcta, de modo que para nueve respuestas incorrectas se esperarían tres conjeturas correctas en otras preguntas. La puntuación corregida de la niña —aquella que de hecho se informa y se compara con las normas existentes— sería entonces de 32, es decir, 35 menos 3. En otras palabras, es probable que conociera 32 respuestas, pero al adivinar en otras 12 aumentó su puntuación en otros 3 puntos.

La corrección para la puntuación que se ejemplifica en el párrafo anterior, se refiere sólo a las respuestas aleatorias, sin conocimientos. El efecto de tal corrección consiste en eliminar las ventajas que de otra manera se concederían a quienes toman riesgos de manera aventurada. En algunos casos, un individuo examinado puede eliminar una o dos de las alternativas, con lo cual aumenta las probabilidades de adivinar correctamente entre las opciones restantes. En esta situación puede ser adecuado que la persona trate de adivinar la respuesta.

El que una conjetura basada en conocimientos conceda en realidad una ventaja a la persona examinada, depende en parte de la diabólica habilidad de quien escribe el reactivo. Traxler (1951) señala que:

> En efecto, el redactor del reactivo intenta hacer que cada respuesta incorrecta parezca tan plausible que todos los examinandos que no poseen la habilidad o capacidad deseada seleccionarán la respuesta incorrecta. En otras palabras, la meta del redactor del reactivo consiste en lograr que todas o casi todas las conjeturas consideradas sean conjeturas erróneas.

Un redactor hábil de reactivos puede diseñar preguntas de modo que la alternativa correcta sea totalmente contraria a la intuición y las alternativas incorrectas tengan una cualidad persuasivamente atractiva. Para estos reactivos, una conjetura basada en conocimientos es casi siempre incorrecta.

Sin embargo, muchos creadores de pruebas le aconsejan ahora a los sujetos que realicen conjeturas basadas en sus conocimientos, pero advierten en contra de las conjeturas aleatorias. Por ejemplo, una reciente edición del manual de preparación para la prueba SAT, *Taking the SAT*, aconseja:

> Debido a la manera en que se califica la prueba, es poco probable que las conjeturas fortuitas o aleatorias para las preguntas de las que usted no sabe nada cambien su calificación. Cuando usted sabe que se pueden eliminar una o más opciones, adivinar la respuesta entre las opciones restantes puede constituir una ventaja a su favor.

Ya sea que una prueba grupal utilice o no una corrección para la puntuación, el punto importante reside en enfatizar en este contexto que

quien aplica la prueba debe seguir el procedimiento estandarizado y nunca ofrecer consejo adicional acerca de adivinar las respuestas. En las pruebas grupales, las desviaciones con respecto al manual de instrucciones son simplemente inaceptables.

INFLUENCIA DEL EXAMINADOR

LA IMPORTANCIA DEL *RAPPORT*

Los editores de pruebas instan a los examinadores a establecer un ***rapport*** —una atmósfera cómoda y cálida que sirva para motivar a los examinandos y que eduzca la cooperación. Dar lugar a un ambiente cordial para la prueba es un aspecto crucial de una prueba válida. Un examinador que no establece *rapport* puede provocar que una persona reaccione con ansiedad, falta de cooperación de naturaleza pasivo-agresiva u hostilidad abierta. El fracaso para establecer el *rapport* distorsiona los datos de la prueba: se subestima la capacidad y se hace un juicio inadecuado de la personalidad.

El *rapport* es de particular importancia en las pruebas individuales y en particular cuando se evalúa a niños. Wechsler (1974) ha señalado que el establecimiento del *rapport* crea fuertes demandas sobre las habilidades clínicas del examinador:

> Debe lograr que el niño se sienta cómodo, mantenerlo interesado en las tareas en cuestión y alentarlo a realizar su mejor esfuerzo. No existe una fórmula mágica para "ganarse" al niño; las aproximaciones que tienen éxito con algunos niños pueden enfadar a otros. Con experiencia, el examinador desarrollará una perceptividad que le permitirá establecer relaciones cordiales con los niños y adaptarse a las necesidades específicas de cada uno. Las sugerencias generales que aparecen a continuación se ofrecen para ayudar al examinador en sus esfuerzos.
>
> Para lograr que el niño se sienta cómodo con su entorno, el examinador podría hacer que participe en alguna conversación informal, antes de adentrarse en el asunto más serio de aplicar la prueba. Con frecuencia, hablar con él acerca de sus pasatiempos o intereses es una buena manera de romper el hielo, aunque en el caso de un niño tímido, puede ser mejor alentarlo a que hable sobre algún aspecto concreto del ambiente —un cuadro en la pared, un animal en el salón de clases o un libro o juguete (que no sea parte del material de prueba) en la habitación de examen. En general, este periodo introductorio no requiere más de 5 a 10 minutos, aunque la prueba no debe comenzar hasta que el niño parezca lo suficientemente relajado como para dar su máximo esfuerzo.

Un estudio de Gregory, Lehman y Mohan (1976) ilustra la importancia de establecer *rapport* cuando se examina a los niños. Estos investigadores buscaron determinar los efectos del bajo nivel de exposición al plomo sobre el CI al aplicar el WISC a 193 niños que vivían cerca de una fundición de plomo. Los niños se asignaron a cinco diferentes estudiantes de posgrado que aplicaron las pruebas, con base en una rotación casi aleatoria en la que el examinador que estuviera libre atendía al primer niño que llegara. Los grupos de niños sometidos a prueba por cada uno de los cinco psicómetras no difería en el promedio de edad, exposición al plomo o clase social. Lo que es más, las magnitudes de las muestras eran sustanciales, ya que su número iba de los 30 a los 45 niños. De aquí que el promedio de los CI examinados dentro de los cinco grupos debería haber sido muy similar.

Sin embargo, las diferencias entre los CI examinados en los cinco grupos fueron penosamente grandes, con puntuaciones promedio que variaban hasta 14 puntos. Clasificadas de menor a mayor, las puntuaciones promedio de los cinco grupos fueron 90, 94, 95, 96 y 104. El examinador cuyos sujetos obtuvieron en promedio un CI de 90 era muy formal, preciso, frío y apresurado. De hecho examinó con mucho a la mayoría de los sujetos (45, en comparación con 37 del siguiente examinador más prolífico) y, en general, terminaba mucho antes con cada niño. En el otro extremo se encontraba el examinador, cuyos sujetos obtuvieron un CI promedio de 104. Fue mucho más allá del buen *rapport*, hasta llegar al grado de ofrecer apoyo y aliento que rayaba en conducir a los sujetos a la respuesta correcta. Por

ejemplo, en Diseño con Cubos instó a un niño a "Vamos, coloca los cubos de las esquinas y sigue a partir de allí".

Por tanto, los examinadores pueden diferir en cuanto a sus capacidades para establecer *rapport*. Es probable que los examinadores fríos obtengan menos cooperación de sus sujetos, lo cual tiene como consecuencia que se reduzca el desempeño en una prueba de capacidad o que se obtengan resultados distorsionados o defensivos en pruebas de personalidad. Los examinadores demasiado solícitos pueden errar en la dirección opuesta, dando señales sutiles (y en ocasiones más que obvias) de las respuestas correctas. Se deben evitar ambos extremos.

SEXO, EXPERIENCIA Y RAZA DEL EXAMINADOR

Un amplio conjunto de investigaciones ha buscado determinar si ciertas características del examinador causan el aumento o la disminución de las puntuaciones de las personas evaluadas con pruebas de capacidad. Por ejemplo, ¿tiene importancia que el examinador sea varón o mujer? ¿Que tenga experiencia o que sea un novato? ¿Que sea de la misma raza o de una diferente a la del sujeto? Nos abstendremos de revisar estos estudios —con unas cuantas excepciones— por una simple razón: los resultados son contradictorios y, por tanto, no son concluyentes. La mayoría de los estudios encuentran que el sexo, la experiencia y la raza del examinador causan poca o ninguna diferencia. Lo que es más, los pocos estudios que informan de un fuerte efecto en una dirección (p. ej., las mujeres examinadoras obtienen puntuaciones más altas de CI de sus sujetos), se contradicen por otros estudios que muestran la tendencia contraria. El lector interesado puede consultar a Sattler (1988) para un análisis y un extenso listado de referencias.

Sin embargo, sería incorrecto concluir que el sexo, la experiencia o la raza del examinador nunca afectan las puntuaciones de prueba. En casos aislados, bien podría ser que una característica particular de un examinador tuviera un fuerte efecto sobre las puntuaciones de prueba de la persona evaluada. Por ejemplo, Terrell, Terrell y Taylor (1981) demostraron de manera ingeniosa que la raza del examinador tiene una potente interacción con el nivel de confianza de los individuos afroestadounidenses sometidos a una prueba de CI. Estos investigadores identificaron a estudiantes universitarios afroestadounidenses con niveles alto y bajo de desconfianza hacia las personas de raza blanca; entonces, un examinador de raza blanca le aplicó el WAIS a una mitad de cada grupo, mientras que a la otra mitad la evaluó un examinador afroestadounidense. El grupo con alto nivel de desconfianza al que examinó un psicómetra afroestadounidense tuvo una puntuación significativamente mayor, que el grupo con alto nivel de desconfianza evaluado por un examinador de raza blanca (CI promedio de 96 contra 86, respectivamente). Además, el grupo con bajo nivel de desconfianza examinado por un psicómetra de raza blanca tuvo puntuaciones un tanto mayores que el grupo con bajo nivel de desconfianza evaluado por el examinador afroestadounidense (CI promedio de 97 contra 92, respectivamente). En suma, los autores concluyeron que los afroestadounidenses desconfiados tienen un desempeño deficiente cuando los someten a prueba examinadores de raza blanca. Los datos que tienen que ver con este tipo de efecto racial son pocos y seguramente queda espacio para realizar investigaciones adicionales.

ANTECEDENTES Y MOTIVACIÓN DE LA PERSONA EXAMINADA

Los examinandos difieren no sólo en las características que los examinadores desean evaluar, sino también en otros aspectos ajenos que podrían alterar los resultados de prueba. Por ejemplo, un individuo inteligente podría tener un desempeño deficiente en una prueba cronometrada de la capacidad debido a ansiedad ante ésta; un asesino cuerdo podría tener el propósito de parecer mentalmente enfermo en un

inventario de personalidad con la finalidad de evitar una acusación; un estudiante con capacidad promedio podría prepararse para tener un mejor desempeño en una prueba de aptitud. Algunos sujetos carecen totalmente de motivación y no les interesa si obtienen un buen resultado en pruebas psicológicas. En todos estos casos, los resultados de prueba pueden ser imprecisos debido a los efectos penetrantes y causantes de distorsión de ciertas características del examinando como la ansiedad, simulación, entrenamiento o antecedentes culturales.

ANSIEDAD ANTE LA PRUEBA

La **ansiedad ante la prueba** se refiere a aquellas respuestas fenomenológicas, fisiológicas y conductuales que acompañan a la preocupación acerca del posible fracaso en una prueba. No hay duda de que los sujetos experimentan diferentes niveles de ansiedad ante la prueba, que van desde una actitud despreocupada, hasta un temor incapacitante ante el prospecto de que se les someta a prueba.

Se han desarrollado varios cuestionarios de verdadero-falso para evaluar las diferencias individuales en ansiedad ante la prueba (p. ej., Sarason, 1980; Morris, Davis y Hutchings, 1981). A continuación se enumeran reactivos característicos y la dirección de su respuesta (V para verdadero, F para Falso):

(V) Cuando realizo un examen importante, sudo en gran cantidad.
(V) Me paralizo cuando presento pruebas de inteligencia o exámenes escolares.
(F) En realidad no comprendo por qué las personas se angustian tanto acerca de las pruebas.
(V) Me aterran los cursos donde al instructor le gusta hacer exámenes "sorpresa".

Un amplio cuerpo de investigación ha confirmado el concepto de sentido común de que la ansiedad ante la prueba se correlaciona en sentido negativo con el rendimiento escolar, las puntuaciones de pruebas de aptitud y las medidas de inteligencia (Naveh-Benjamin, McKeachie y Lin, 1987; McKeachie, 1984). Sin embargo, la interpretación de estos descubrimientos de correlación no es simple. Una posibilidad es que los estudiantes desarrollen ansiedad ante la prueba, debido a antecedentes de tener un mal desempeño en los exámenes. Es decir, las disminuciones en desempeño pueden anteceder y causar dicha ansiedad. En apoyo a este punto de vista, Paulman y Kennelly (1984) encontraron que —con independencia de su ansiedad— muchos estudiantes con ansiedad ante la prueba también tenían ineficiencias en presentación de exámenes en ambientes académicos. Dichos estudiantes lograban pobres resultados en pruebas, estuvieran ansiosos o no. Lo que es más, Naveh-Benjamin, McKeachie y Lin (1987) determinaron que una gran proporción de estudiantes universitarios con ansiedad ante la prueba tenían malos hábitos de estudio que los predisponían a un pobre desempeño en pruebas. La ansiedad de estos individuos es, en parte, un subproducto de la frustración que han tenido toda la vida con respecto a resultados mediocres en pruebas.

Otras líneas de investigación indican que la ansiedad ante la prueba tiene un efecto perjudicial directo sobre el desempeño en pruebas. Es decir, es probable que esta ansiedad sea tanto una causa como un efecto en la ecuación que la vincula con el desempeño deficiente en pruebas. Considérese el estudio original que realizó Sarason (1961) sobre este tema, quien sometió a prueba a sujetos con alta y baja ansiedad bajo instrucciones neutras o inductoras de ansiedad. Los sujetos eran estudiantes universitarios a los que se les pidió que memorizaran palabras de dos sílabas con un bajo nivel de significado —una tarea difícil. La mitad de los sujetos trabajó bajo instrucciones neutras —simplemente se les dijo que memorizaran las listas. A los sujetos restantes se les dijo que memorizaran las listas y que la tarea era una prueba de inteligencia. Se les instó a tener el mejor desempeño posible. Los dos grupos no difirieron de manera significativa en su desempeño

cuando las instrucciones eran neutras y no amenazantes. Sin embargo, cuando éstas excitaban la ansiedad, los niveles de desempeño de los sujetos con alto grado de ansiedad descendieron de modo notable, dejándolos con una enorme desventaja, en comparación con aquellos que tenían baja ansiedad. Esto indica que los sujetos con ansiedad ante la prueba muestran disminuciones significativas en su desempeño cuando perciben la situación como una prueba. En contraste, los sujetos con bajo nivel de ansiedad sufren un efecto relativamente nulo de tal redefinición simple del contexto.

Las pruebas con estrictos límites de tiempo representan un problema especial para las personas con altos niveles de ansiedad ante la prueba. La presión de tiempo parece exacerbar el grado de amenaza personal, causando reducciones significativas en el desempeño de las personas con ansiedad ante la prueba. Siegman (1956) demostró este punto hace muchos años al comparar los niveles de desempeño de pacientes médicos/psiquiátricos con alta y baja ansiedad en las subpruebas con y sin límite de tiempo del WAIS. El WAIS consiste de 11 subpruebas que incluyen seis subpruebas para las cuales el examinador utiliza un cronómetro, a fin de imponer estrictos límites de tiempo y cinco subpruebas para las cuales el sujeto tiene un tiempo ilimitado de respuesta. Es interesante que los sujetos con alta y baja ansiedad tuvieron una capacidad total idéntica en el WAIS. Sin embargo, cada grupo tuvo un mejor desempeño en la dirección prevista en los diferentes tipos de subpruebas. En particular, los sujetos con bajo nivel de ansiedad superaron a aquéllos con ansiedad elevada en las subpruebas con límite de tiempo, mientras que se observó el patrón inverso en las subpruebas sin límite de tiempo (figura 2–4).

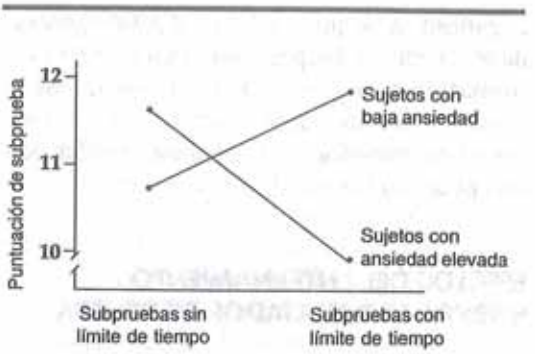


Figura 2–4. Influencia de los límites de tiempo y del nivel de ansiedad en los resultados de subpruebas del WAIS.
Fuente: Basado en datos de Siegman, A. W. (1956). The effect of manifest anxiety on a concept formation task, a nondirected learning task, and on timed and untimed intelligence tests. *Journal of Consulting Psychology, 20,* 176-178.

MOTIVACIÓN PARA EL ENGAÑO

Los resultados de prueba pueden ser imprecisos si la persona tiene razones para desempeñarse de una manera inadecuada o no representativa. El falseamiento abierto de los resultados de prueba es poco común, pero llega a suceder. Una pequeña fracción de personas que buscan beneficiarse de la rehabilitación o de instituciones sociales "se fingirán malos" en pruebas de personalidad o de capacidad. En ocasiones, las personas que anticipan una acusación criminal fingirán enfermedad mental en pruebas de personalidad. Considérese el caso del cliente de psicoterapia que contestó una prueba de personalidad por instrucciones de su terapeuta. El terapeuta deseaba una evaluación precisa de la depresión aparentemente leve del cliente. Los resultados fueron ambiguos, lo cual indica ya sea un grado monumental de trastorno psicológico o un intento consciente de exagerar los síntomas. Dos semanas después, el terapeuta descubre de manera inadvertida que el cliente está a punto de ser acusado de abuso sexual infantil. En apariencia, había falseado los resultados de prueba anticipando que pronto se harían cargos legales en su contra. En parte, había planeado defenderse aduciendo que la enfermedad mental era una factor atenuante para su conducta.

En la mayoría de los casos, un psicómetra bien entrenado puede detectar el falseamiento consciente al hacer dos preguntas: 1) ¿El cliente tiene motivación para tener un desempeño

engañoso en las pruebas? 2) ¿El patrón general de resultados es sospechoso a la luz de otra información conocida acerca del cliente? Si la respuesta para ambas preguntas es "Sí", entonces el examinador haría bien en abordar con escepticismo los resultados de prueba.

EFECTOS DEL ENTRENAMIENTO SOBRE LOS RESULTADOS DE PRUEBA

Psicólogos y educadores han estudiado ampliamente la influencia del entrenamiento sobre los resultados de prueba. El **entrenamiento** puede incluir varios componentes: práctica adicional en materiales similares a la prueba, revisión de conceptos fundamentales que es probable que cubra la prueba y consejo sobre las estrategias óptimas de presentación de prueba. Aquí sólo se pueden mencionar unos cuantos puntos sobresalientes, con un énfasis particular sobre las pruebas de aptitud. Los lectores que deseen conocer más detalles pueden consultar a Anastasi (1981) y Bond (1989).

No hay duda que el entrenamiento puede aumentar significativamente las puntuaciones en ciertos tipos de pruebas de aptitud que contienen tipos de reactivos de prueba que pueden "entrenarse". Por ejemplo, Powers y Swinton (1984) enviaron por correo varios conjuntos de materiales de preparación de prueba a muestras aleatorias de candidatos para el Examen de Selección para el Posgrado (GRE, del inglés *Graduate Record Examination*), aproximadamente cinco semanas antes de la aplicación de la prueba.* Los materiales de preparación para la prueba incluían pruebas adicionales de práctica, explicaciones para las preguntas de las pruebas de práctica y pistas o estrategias para responder diferentes tipos de reactivos. Incidentalmente, las calificaciones de prueba en el GRE van, por lo general, de 200 a 800, con un promedio nacional de aproximadamente 500. La prueba general incluye tres secciones: verbal, cuantitativa y analítica. A través de la comparación del desempeño de los sujetos experimentales con los sujetos control que no recibieron materiales complementarios de preparación de prueba, Powers y Swinton (1984) pudieron deducir que cuatro horas de preparación de autoadiestramiento producían como dividendo un aumento de 53 puntos en la escala de Prueba Analítica del GRE. Este descubrimiento fue comparable con un aumento de 66 puntos en una intervención a través de un instructor, la cual implicaba siete horas de contacto directo (Swinton y Powers, 1983).

Sin embargo, los efectos observados en estos estudios fueron sumamente específicos. Los materiales especiales de preparación para los reactivos analíticos actuaron sólo sobre las puntuaciones analíticas, no sobre las verbales o cuantitativas. De hecho, los efectos se restringieron sólo a los tipos particulares de reactivos analíticos que eran **entrenables** —aquellos que implicaban análisis de explicaciones y diagramas lógicos. El desempeño en otros tipos de reactivos analíticos —por ejemplo, razonamiento analítico— no se vio afectado.

El entrenamiento representa un grave problema para quienes desarrollan pruebas, debido a que puede inflar la puntuación de una persona sin mejorar de manera correspondiente sus capacidades generales en el dominio que se somete a prueba. Si esto ocurre, la puntuación ya no es una representación básica de la capacidad del examinando. Así, el entrenamiento puede invalidar las normas existentes. Con toda seguridad, el entrenamiento hace surgir el fantasma de los privilegios en cuanto a preparación de pruebas para los estudiantes ricos o astutos, a expensas de aquellos pobres o desinformados. Después de todo, abundan las escuelas privadas de entrenamiento que pueden cobrar miles de dólares por un día o dos de preparación para los exámenes de aptitud.

Por fortuna, existe una solución parcial para el problema del entrenamiento: poner a disposición de todos el entrenamiento a través del autoadiestramiento. La mayoría de los principales programas de prueba, incluyendo al GRE,

* Los comités de aceptación de muchas escuelas de posgrado requieren el GRE. Durante el periodo en que Powers y Swinton (1984) estudiaron el GRE, este examen consistía de las partes verbal, cuantitativa y analítica.

proporcionan ahora muestras de los materiales de prueba, de modo que los individuos que presenten el examen puedan familiarizarse con la naturaleza de las preguntas. Por supuesto, esto no garantiza que todos utilizarán los materiales, pero al menos la oportunidad está disponible de manera inmediata sin ningún costo. Los estudiantes que se niegan a inspeccionar los materiales de familiarización con la prueba lo hacen bajo su cuenta y riesgo.

ASPECTOS SOBRE CALIFICACIÓN

En general, las pruebas grupales emplean un formato de opción múltiple en el cual el examinando marca con un lápiz las respuestas en una hoja separada de respuestas, la que se califica entonces a través de una máquina con total objetividad y precisión. En consecuencia, es poco común que las pruebas grupales presenten una oportunidad para el error humano en la calificación. Sin embargo, los errores de calificación pueden colarse en las pruebas grupales calificadas por medios mecánicos, si las personas examinadas no oscurecen lo suficiente las áreas de respuesta con un lápiz de punto suave o si dejan marcas no relacionadas en la hoja de respuesta. Para protegerse de esta fuente de error, los examinadores deben revisar cada hoja de respuesta y corregir cualquier irregularidad en las marcas con lápiz, antes de someter los materiales a la calificación mecánica.

Los errores de calificación ocurren con las pruebas individuales en las que el examinador debe hacer juicios de calificación, sumar columnas de números y consultar tablas para convertir las puntuaciones brutas en CI o en otras estadísticas de resumen. Contrario a la creencia popular, los juicios del examinador con respecto a la calificación —por ejemplo, el que una cierta respuesta en una prueba de CI merezca 1 o 2 puntos de puntuación bruta— casi nunca representan una fuente de error de calificación. Se ilustrará esta situación con una prueba individual del CI, pero ello se aplica también a la mayoría de las demás pruebas individuales. Incluso cuando los examinadores se equivocan de manera flagrante en una dirección conservadora, concediendo de manera consistente a un sujeto un crédito menor del que merece por respuestas ambiguas o limítrofes, el efecto neto sobre el CI Total es mínimo, quizás en uno o dos puntos cuando más. La razón para que los errores de juicio raramente signifiquen una diferencia seria es sencilla: los criterios de calificación de la mayoría de las pruebas se expresan con tanto detalle que en raras ocasiones se demanda "un juicio" por parte del examinador.

Los errores de calificación de índole administrativa son totalmente otra situación. Estos tipos de errores ocurren con mucha mayor frecuencia de lo que los psicólogos desearían admitir alguna vez. Como se verá, los errores administrativos de calificación pueden tener efectos desastrosos.

Ryan, Prifitera y Powers (1983) le pidieron a 19 psicólogos y a 20 estudiantes de posgrado que calificaran los protocolos del WAIS-R de dos clientes de orientación vocacional. Para un cliente cuyo CI Total era de 110, los psicólogos practicantes calcularon puntuaciones que iban de 107 a 115 y los estudiantes de posgrado obtuvieron puntuaciones que iban de 108 a 117. Las variaciones en puntuación se debieron en gran medida a **errores administrativos de calificación**, no a diferencias de juicio en cuanto al crédito adecuado para respuestas ambiguas o limítrofes. Gregory (1987) ha ilustrado en mucho el mismo punto con un grupo de estudiantes avanzados de posgrado que cometieron errores que abarcaron hasta 30 puntos al calificar un protocolo de una prueba estándar de CI, de Wechsler. Por tanto, se puede conjeturar que los errores de calificación sí ocurren con frecuencia y comprometen gravemente la precisión de la evaluación del CI y de otras formas de pruebas psicológicas. Este problema no desaparece tan sólo con un aumento en la experiencia. La única manera de evitar los errores administrativos de calificación consiste en exponer qué tan extendido está el problema y alentar a los examinadores a ejercer un gran cuidado en la calificación de los protocolos.

RESUMEN

1. Los procedimientos estandarizados de examen son esenciales para una prueba válida. El uso de procedimientos no estandarizados puede alterar el significado de los resultados de prueba, lo cual los vuelve inválidos y engañosos.
2. Sin embargo, la flexibilidad en los procedimientos de prueba es apropiada cuando es razonada y deliberada. Para determinar si es aceptable un cambio flexible en los procedimientos de prueba, el examinador debe suponer cómo es más probable que se haya aplicado la prueba a la muestra normativa.
3. En las pruebas individuales, es deseable que el examinador se familiarice en gran medida con los materiales de prueba. Es necesario que se ensaye con la prueba, de modo que el examinador pueda anticipar las respuestas apropiadas ante las numerosas contingencias que ocurren durante la aplicación de la misma.
4. Otro ingrediente importante en una prueba válida es la sensibilidad a las discapacidades de la persona examinada. Cuando no se reconocen las discapacidades, pueden ocurrir graves errores en la interpretación de la prueba; por ejemplo, a una persona con sordera se le puede diagnosticar de manera equivocada como si tuviera retraso mental.
5. En la aplicación de pruebas grupales, los examinadores deben adherirse de modo estricto a las instrucciones orales y a los límites de tiempo definidos. Además, las condiciones de prueba deben ser apropiadas; por ejemplo, iluminación adecuada y ruido mínimo.
6. En especial cuando se trata de la aplicación de pruebas individuales, se insta a los examinadores a establecer el *rapport*. En pruebas, el *rapport* es una atmósfera cómoda y cálida que sirve para motivar a los examinandos y para facilitar la cooperación.
7. Por el contrario de la expectativa popular, la mayoría de los estudios encuentran que el sexo, la experiencia y la raza del examinador tienen poco efecto sobre los resultados de una prueba psicológica. Sin embargo, pueden haber casos especiales en los que las interacciones examinador-examinando producen efectos perjudiciales sobre las puntuaciones de prueba.
8. La ansiedad ante la prueba se refiere a aquellas respuestas fenomenológicas, fisiológicas y conductuales que acompañan a la preocupación sobre un posible fracaso en una prueba. Se ha mostrado que la ansiedad ante la prueba se correlaciona en sentido negativo con el rendimiento escolar, las puntuaciones en pruebas de aptitud, las medidas de inteligencia y con el desempeño en pruebas con límite de tiempo.
9. El falseamiento de los resultados de prueba es poco común, pero llega a ocurrir. En la mayoría de los casos, un examinador bien entrenado puede detectar el falseamiento consciente al cuestionar si el cliente tiene motivación para tener un desempeño engañoso en las pruebas o si el patrón general de resultados de prueba es sospechoso a la luz de otras informaciones.
10. El entrenamiento puede mejorar las calificaciones de prueba en ciertos tipos de reactivos entrenables, como la porción analítica del Examen de Selección para el Posgrado (GRE). La práctica con reactivos de muestra y las pistas o estrategias de aprendizaje pueden elevar la puntuación de subprueba de la GRE en 50 a 60 puntos en la escala de 800 puntos.
11. En la calificación de pruebas individuales, las diferencias en el juicio personal —por ejemplo, asignar a una respuesta de prueba un punto en lugar de dos en puntuación bruta— raramente influyen sobre la puntuación total de prueba en algún grado considerable; por ejemplo, 1 o 2 puntos del CI Total, cuando más.

12. Los errores administrativos de calificación —p. ej., sumar de manera incorrecta columnas de puntuaciones o consultar la tabla inadecuada de referencia— representan un serio problema en las pruebas psicológicas individuales. Estos tipos de errores pueden causar que las puntuaciones de prueba sean obviamente imprecisas.

TÉRMINOS Y CONCEPTOS CLAVE

Corrección de conjeturas p. 59
Rapport p. 60
Ansiedad ante la prueba p. 62
Entrenamiento p. 64
Errores administrativos de calificación p. 65

CAPÍTULO 3

Normas y confiabilidad

TEMA 3A
Normas y estandarización

Puntuaciones naturales
Ejemplo de caso 3–1. Pruebas anticuadas y normas obsoletas
Conceptos estadísticos esenciales
Transformación de las puntuaciones naturales
Selección de un grupo normativo
Pruebas referidas a criterio
Resumen
Términos y conceptos clave

Este capítulo se ocupa de dos conceptos básicos necesarios para facilitar la interpretación del examinador acerca de las puntuaciones de prueba: las normas y la confiabilidad. En la mayor parte de los casos, las puntuaciones en pruebas psicológicas se interpretan con referencia a normas basadas en la distribución de las puntuaciones obtenidas por una muestra representativa de personas examinadas. En el tema 3A, Normas y estandarización, se revisa el proceso de estandarización de una prueba contra un grupo normativo apropiado, de modo que los usuarios de la prueba puedan darle sentido a las puntuaciones individuales en esa prueba. Dado que la utilidad de una puntuación de prueba también se determina por la consistencia o replicabilidad de los resultados de la misma, se introducen los aspectos esenciales de la teoría sobre confiabilidad y medición en el tema 3B, Conceptos de confiabilidad. El siguiente capítulo se origina lógicamente del material presentado aquí e investiga los complejos problemas

sobre validez —¿una prueba mide lo que se supone que mida? Primero, se comienza con los problemas más directos acerca del establecimiento de un marco de referencia comparativo (normas) y la determinación de la consistencia o replicabilidad de los resultados de prueba (confiabilidad).

Por lo general, el resultado inicial de las pruebas es una puntuación natural como el número total de afirmaciones de personalidad refrendadas en una dirección en particular o el número total de problemas resueltos correctamente, quizá con la adición de puntos bonificados por las soluciones rápidas. En la mayor parte de los casos, esta puntuación inicial es inútil en sí misma. Para que los resultados de prueba tengan significado, los examinadores deben ser capaces de convertir la puntuación inicial a alguna forma de puntuación derivada que se base en la comparación con un grupo normativo o de estandarización. Gran parte de las pruebas se interpreta al comparar los resultados individuales con el desempeño del grupo normativo; las pruebas referidas a criterio, que se analizan más adelante, constituyen una excepción.

Un **grupo normativo** consiste en una muestra de personas examinadas que son representativas de la población hacia la cual se dirige la prueba. Considérese una prueba del conocimiento de vocabulario, diseñada para utilizarse con futuros estudiantes universitarios de primer grado. En este caso, podrían recolectarse los resultados del desempeño de una muestra grande, heterogénea y nacional de dichas personas, con propósitos de estandarización. El objetivo esencial de la estandarización de una prueba consiste en determinar la distribución de las puntuaciones naturales en un grupo normativo, de modo que quienes desarrollan la prueba tengan la posibilidad de publicar las puntuaciones derivadas conocidas como normas. Como se analizará más adelante, las normas se encuentran en muchas variedades; por ejemplo, rangos percentiles, equivalentes por edad, equivalentes de grado o puntuaciones estándar. En general, las normas indican la posición que tiene un individuo dentro de la prueba, en relación con el desempeño de otras personas de la misma edad, grado escolar, sexo y otras variables.

Para ser efectivas, las normas deben obtenerse con gran cuidado y construirse según los preceptos ya conocidos que se analizan después; lo que es más, pueden volverse anticuadas en sólo unos cuantos años, de modo que la regla, más que la excepción, debe ser el establecimiento periódico de nuevas normas (ejemplo de caso 3–1). El tema de las normas se enfoca de manera indirecta, ya que primero se le proporciona al lector un análisis sobre las puntuaciones naturales y después se revisan los conceptos estadísticos esenciales para una comprensión de las normas.

PUNTUACIONES NATURALES

El nivel más básico de información proporcionada por una prueba psicológica es la **puntuación natural**. Por ejemplo, en una prueba de personalidad, con frecuencia la puntuación natural es el número de preguntas respondidas en la dirección codificada para una escala específica. En las pruebas de capacidad, por lo común la puntuación natural se integra del número de problemas respondidos de manera correcta, al que con frecuencia se le suman los puntos bonificados por el desempeño rápido. Así, el resultado inicial de la prueba es casi siempre una suma numérica, como 17 de 44 reactivos totales respondidos en la dirección codificada en una escala de depresión, o 29 de 55 puntos de puntuación natural obtenidos en la subescala de Diseño con Cubos de una prueba de inteligencia.

Sin embargo, debería ser obvio para el lector que las puntuaciones naturales, por sí solas carecen absolutamente de significado. Por ejemplo, ¿de qué sirve conocer que una persona resolvió de manera correcta 12 de 20 preguntas de razonamiento abstracto? ¿Qué significado tiene que un examinando haya respondido en la dirección codificada a 19 de 33 preguntas de verdadero-falso de una escala de disposición psicológica?

EJEMPLO DE CASO 3–1
PRUEBAS ANTICUADAS Y NORMAS OBSOLETAS

Un psicólogo consultor trabajaba en un sistema de educación secundaria que alardeaba de un respetado programa para niños dotados y talentosos. El ingreso a este programa podía obtenerse de diversas maneras, incluyendo la documentación por parte de un psicólogo de que el solicitante tenía un CI de 130 o mayor. Los funcionarios de la escuela se sorprendieron cuando el número de estudiantes aceptados aumentó de manera abrupta a inicios del decenio de 1980. La mayoría de los recién encontrados estudiantes con alto CI habían pasado por una prueba con un psicólogo educativo, quien era una persona honrada y competente que no tenía una aparente motivación para falsear los resultados. El director de la escuela comentó con el psicólogo consultor que la comunidad parecía estar fomentando una proporción cada vez mayor de estudiantes dotados y talentosos. Cuando se le preguntó qué pruebas estaban utilizando para la selección al programa, el director respondió que el WAIS (Wechsler, 1955) era el instrumento para los estudiantes mayores.

El consultor señaló que ésta había sido una buena prueba en su tiempo, pero que ya había pasado por una revisión (el WAIS-R se publicó en 1981) y que sus normas anteriores ya no eran pertinentes. De hecho, el WAIS es mucho más fácil que su revisión (Kaufman, 1990; Mishra y Brown, 1983) y tiende a producir un CI que es ¡hasta 8 puntos más alto! No es sorprendente que tantos estudiantes hayan obtenido un CI superior. El efecto de utilizar una prueba obsoleta era que, en promedio, se sumaban 8 puntos al CI real de cada solicitante. Al escuchar estos datos, el director determinó comprar una copia del WAIS-R, de modo que el psicólogo de la escuela pudiera tener un instrumento actualizado, con normas apropiadas, para evaluar a los solicitantes del programa especial.

Incluso es difícil pensar sobre dichas preguntas sin recurrir a comparaciones de una variedad u otra. Se desea saber la manera en que otras personas han respondido a estas pruebas, si las puntuaciones observadas son altas o bajas en comparación con un grupo representativo de sujetos. En el caso de pruebas de capacidad, se tiene curiosidad de saber si las preguntas fueron fáciles o difíciles, en especial en relación con la edad del sujeto.

De hecho, parece casi trivial que una puntuación natural adquiera significado principalmente en relación con las normas, un marco de referencia establecido de manera independiente que se deriva de una muestra de estandarización. Posteriormente se ampliará más acerca de la derivación y el uso de las normas. Por ahora bastará con saber que las normas se establecen de modo empírico, a través de la aplicación de la prueba a una muestra grande y representativa de personas. Después se compara la puntuación del examinado con la distribución de puntuaciones obtenidas por la muestra de estandarización. Así, se determina, a partir de las normas, si una puntuación obtenida es baja, promedio o alta.

La gran parte de las pruebas psicológicas se interpreta a través de la consulta de normas; como ya se señaló, estos instrumentos se denominan **pruebas referidas a la norma**. Sin embargo, se recuerda al lector que existen otros tipos de instrumentos. En particular, las pruebas referidas a criterio ayudan a determinar si una persona puede alcanzar un criterio objetivamente definido, como la suma de pares de números de dos dígitos con una precisión de 97%. En el caso de las pruebas referidas a criterio, las normas no son esenciales. Al final de este tema se explicarán con más detalle estas pruebas.

Existen diferentes tipos de normas, pero éstas tienen una característica en común: cada una incorpora un resumen estadístico de un enorme conjunto de puntuaciones. Así, para comprenderlas, el lector necesita dominar la estadística descriptiva elemental. En este punto se hace un modesto paréntesis para revisar los conceptos estadísticos esenciales.

CONCEPTOS ESTADÍSTICOS ESENCIALES

Supóngase por el momento que se tiene acceso a una prueba de vocabulario de alto nivel, apropiada para examinar las habilidades verbales de profesores universitarios y otros profesionales (Gregory y Gernert, 1990). La prueba es un cuestionario de opción múltiple con 30 palabras difíciles como **firmamento, paradisiaco** y **melifluo**. Una profesora toma la prueba y elige la alternativa correcta para 17 de las 30 palabras. Pregunta cómo se compara su calificación con otras personas con el mismo nivel académico. ¿Cómo podría responderse a su pregunta?

Una manera de responder a la interrogante consistiría en darle una lista de las puntuaciones naturales de la muestra preliminar de estandarización con 100 profesores representativos de su universidad (cuadro 3–1). Sin embargo, incluso con esta muestra normativa relativamente pequeña (lo típico son miles de individuos), la lista de puntuaciones de prueba es un despliegue excesivo.

Cuando se nos confronta con un conjunto de datos cuantitativos, la tendencia humana natural es la de resumir, condensar y organizar dichos datos en patrones significativos. Por ejemplo, en la evaluación del significado de la puntuación de vocabulario de la profesora, el lector podría calcular la puntuación promedio de toda la muestra o establecer la posición relativa de la puntuación de la profesora (17 respuestas correctas) entre los 100 datos encontrados en el cuadro 3–1. En las siguientes secciones se revisarán éstos y otros enfoques para la organización y resumen de datos cuantitativos.

Cuadro 3–1. Puntuaciones naturales de 100 profesores en una prueba de vocabulario con 30 reactivos

6,	10,	16,	16,	17,	14,	19,	14,	16,	15
17,	17,	19,	20,	20,	22,	17,	24,	14,	25
13,	20,	11,	20,	21,	11,	20,	16,	18,	12
13,	7,	20,	27,	21,	7,	15,	18,	18,	25
20,	27,	28,	13,	21,	17,	12,	18,	12,	15
9,	24,	25,	9,	17,	17,	9,	19,	24,	15
20,	21,	22,	12,	21,	12,	19,	19,	23,	16
8,	12,	12,	17,	13,	19,	13,	11,	16,	16
7,	19,	14,	17,	19,	14,	18,	15,	15,	15
14,	14,	17,	18,	18,	22,	11,	15,	13,	9

Fuente: Basado en datos de Gregory, R. J. y Gernert, C. H. (1990). *Age trends for fluid and crystallized intelligence in an able subpopulation.* Manuscrito sin publicar.

DISTRIBUCIÓN DE FRECUENCIAS

Una manera simple y útil de resumir los datos consiste en tabular una **distribución de frecuencias** (cuadro 3–2), la cual se prepara al especificar un pequeño número de intervalos de clase de igual tamaño y después determinar cuántas puntuaciones caen dentro de cada intervalo. La suma de las frecuencias de todos los intervalos será igual a N, el número total de

Cuadro 3–2. Distribución de frecuencia de las puntuaciones de 100 profesores en una prueba de vocabulario

Intervalo de clase	Frecuencia
4–6	1
7–9	8
10–12	12
13–15	21
16–18	24
19–21	21
22–24	7
25–27	5
28–30	1
	$N = 100$

puntuaciones en la muestra. No existe una regla simple para determinar el tamaño de los intervalos; éste, obviamente, depende del número de intervalos deseado. Es común que la distribución de frecuencias tenga entre 5 y 15 intervalos de clase. En el caso del cuadro 3–2 existen 9 intervalos de clase con 3 puntuaciones cada uno. El cuadro indica que un profesor obtuvo una calificación de 4, 5 o 6; 8 profesores obtuvieron 7, 8 o 9, y así sucesivamente.

Un **histograma** proporciona una representación gráfica de la misma información contenida en la distribución de frecuencias (figura 3–1A). El eje horizontal representa las puntuaciones agrupadas en intervalos de clase, mientras el eje vertical representa el número de puntuaciones que caen dentro de cada intervalo de clase. En un histograma, la altura de una columna indica el número de puntuaciones que ocurren dentro de ese intervalo. Un **polígono de frecuencias** es similar a un histograma, excepto que la frecuencia de los intervalos de clase se representa con puntos en lugar de columnas. Después, los puntos independientes se unen por medio de líneas rectas (figura 3–1B).

Las gráficas que se muestran en las figuras 3–1 constituyen resúmenes visuales de los 100 datos en puntuación natural de la muestra de profesores. Además de estos resúmenes visuales, también es posible producir resúmenes numéricos mediante el cálculo de los índices de tendencia central y dispersión.



Figura 3–1A. Histograma que representa las puntuaciones en prueba de vocabulario de 100 profesores.

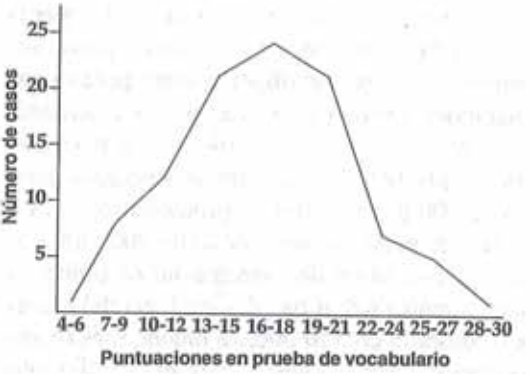


Figura 3–1B. Polígono de frecuencias de las puntuaciones en prueba de vocabulario de 100 profesores.

MEDIDAS DE TENDENCIA CENTRAL

¿Se puede establecer una sola puntuación representativa de las 100 puntuaciones de vocabulario en nuestra muestra? La **media** ($\bar{X}$) o promedio aritmético es una de dichas puntuaciones. Se calcula sumando todas las puntuaciones y dividiéndolas entre N, el número de puntuaciones. Otro índice útil de tendencia central es la **mediana**, la puntuación que se encuentra a la mitad cuando se han ordenado todas las puntuaciones. Si el número de puntuaciones es par, la mediana es el promedio de las dos puntuaciones a la mitad. En cualquier caso, la mediana es el punto que divide en dos la distribución, de modo que la mitad de los casos se encuentren por encima de ella y la mitad por debajo. Por último, la **moda** es simplemente la puntuación que ocurre con mayor frecuencia. Si dos puntuaciones tienen la mayor frecuencia de ocurrencia, se dice que la distribución es bimodal.

La media de las puntuaciones enumeradas en el cuadro 3–1 es 16.8; la mediana y la moda

son 17. En este caso, las tres medidas de tendencia central tienen muy buena concordancia. Sin embargo, no siempre ocurre así. La media es sensible a los valores extremos y puede ser engañosa si una distribución tiene pocas puntuaciones inusualmente altas o bajas. Considérese el caso extremo donde nueve personas ganan $10 000.00 y una décima persona gana $910 000.00. El ingreso promedio para este grupo de personas sería de $100 000; sin embargo, este nivel de ingresos no es típico de nadie dentro del grupo. La mediana del ingreso, colocada en $10 000, es mucho más representativa. Por supuesto, éste es un ejemplo extremo, pero ilustra un punto en general: si una distribución está sesgada (es decir, es asimétrica), la mediana es un mejor índice de la tendencia central que la media.

MEDIDAS DE VARIABILIDAD

Dos o más distribuciones de puntuaciones de prueba pueden tener la misma media y, sin embargo, es posible que difieran en gran medida en el grado de dispersión de las puntuaciones con respecto a la media (figura 3–2). Para describir el grado de dispersión, es necesario un índice estadístico que exprese la variabilidad de las puntuaciones en una distribución.

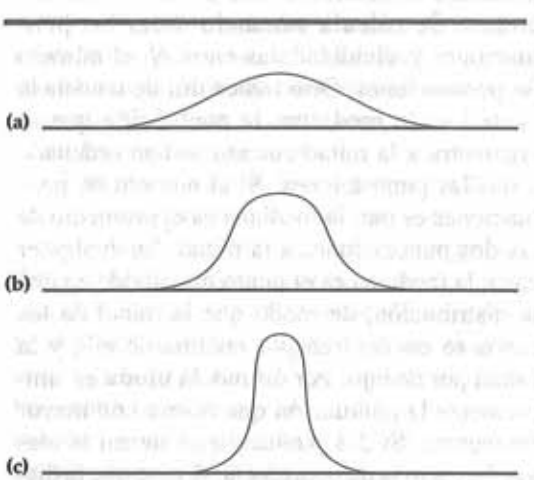


Figura 3–2. Tres distribuciones con medias idénticas, pero con diferente variabilidad.

El índice estadístico de variabilidad que se utiliza con mayor frecuencia en un grupo de puntuaciones es la **desviación estándar**, simbolizada σ y abreviada *DE*. Desde un punto de vista conceptual, el lector necesita saber que la *DE* refleja el grado de dispersión en un grupo de puntuaciones. Si éstas se encuentran agrupadas estrechamente cerca de un valor central, la *DE* es pequeña. De hecho, en el caso extremo donde todas las puntuaciones son idénticas, la *DE* es exactamente cero. A medida que un grupo de puntuaciones se dispersa más, la *DE* se vuelve más grande. Por ejemplo, en la figura 3–2 la distribución **a** tendría la *DE* más grande, la distribución **c**, la más pequeña.

La desviación estándar o σ es, en términos sencillos, la raíz cuadrada de la varianza, denominada σ^2. La fórmula de la **varianza** es

$$\sigma^2 = \frac{\Sigma (X - \overline{X})^2}{(N-1)}$$

donde Σ significa "sumatoria de", X representa cada puntuación individual, $\overline{X}$ es la media de las puntuaciones y N es el número total de puntuaciones. Como lo sugiere el nombre, la varianza es una medida de variabilidad. Sin embargo, en general, los psicólogos prefieren informar la desviación estándar, que se calcula obteniendo la raíz cuadrada de la varianza. Por supuesto, la varianza y la desviación estándar transmiten información intercambiable —una se puede calcular a partir de la otra, al elevar al cuadrado (la desviación estándar para obtener la varianza) u obtener la raíz cuadrada (de la varianza para obtener la desviación estándar). Sin embargo, la desviación estándar es la medida preferida de varianza en las pruebas psicológicas, debido a su relevancia directa para la distribución normal, como se analizará en la siguiente sección.

DISTRIBUCIÓN NORMAL

El polígono de frecuencia que se representa en la figura 3–1B es sumamente irregular en cuanto

a su forma, un hallazgo típico con los datos de la vida real que se basan en muestras pequeñas. ¿Qué sucedería con la forma del polígono de frecuencia si se aumentara la magnitud de la muestra normativa y también se incrementara el número de intervalos de clase al reducir su tamaño? Es posible que, a medida que se añaden nuevos individuos a la muestra, la distribución de puntuaciones se parezca cada vez más a una curva simétrica, definida matemáticamente y en forma de campana, denominada **distribución normal** (figura 3–3).

Los psicólogos prefieren una distribución normal de puntuaciones de prueba, aun cuando muchas otras distribuciones son teóricamente posibles. Por ejemplo, dentro de la esfera de posibilidades se encuentra la distribución rectangular de puntuaciones de prueba –un número igual de resultados en cada intervalo de clase. De hecho, muchos legos podrían incluso preferir una distribución rectangular de puntuaciones de prueba, basados en la premisa equitativa de que las diferencias individuales serían por ello menos pronunciadas. Por ejemplo, una mayor proporción de personas obtendría puntuaciones en el rango superior si las pruebas psicológicas se conformaran a una distribución rectangular de puntuaciones, en vez de a una distribución normal.

Entonces, ¿por qué los psicólogos prefieren una distribución normal de puntuaciones, aun hasta el punto de seleccionar reactivos de prueba que ayuden a producir este tipo de distribución en la muestra de estandarización? Existen varias razones para ello, incluyendo las consideraciones estadísticas y los datos empíricos. Aquí se hará un breve paréntesis para explicar la fascinación psicométrica con las distribuciones normales.

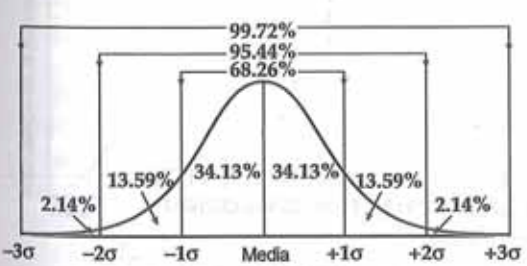


Figura 3–3. Curva normal y porcentaje de casos dentro de ciertos intervalos.

Una razón por la que los psicólogos prefieren las distribuciones normales es que la curva normal tiene características matemáticas útiles que forman la base para varios tipos de investigación estadística. Supóngase que se tiene interés en determinar si los CI promedio de dos grupos de personas fueron significativamente diferentes. Sería apropiado utilizar una estadística inferencial como la prueba *t* para la diferencia entre medias. Sin embargo, muchas estadísticas inferenciales se basan en la suposición de que la población subyacente de puntuaciones se distribuye de manera normal, o muy cercano a ello. Así, a fin de facilitar el uso de estadísticas inferenciales, los psicólogos prefieren que las puntuaciones de prueba en la población normal sigan una distribución normal o casi normal.

Otra base para preferir la distribución normal es su precisión matemática. Dado que la distribución normal se define de manera precisa en términos matemáticos, es posible calcular con gran exactitud el área bajo las diferentes regiones de la curva. Así, una propiedad útil de las distribuciones normales es que el porcentaje de casos que caen dentro de un cierto rango o más allá de un cierto valor se conoce de manera exacta. Por ejemplo, en una distribución normal, tan sólo 2.14% de las puntuaciones excederán la media en dos desviaciones estándar o más (figura 3–3). De la misma manera, se puede determinar que la gran parte de las puntuaciones —más de 68%— caen dentro del rango de una *DE* a partir de la media, en cualquier dirección.

Una tercera base para preferir una distribución normal de las puntuaciones de prueba es que, con frecuencia, la curva normal surge de manera espontánea en la naturaleza. De hecho, los primeros investigadores se impresionaron tanto con la universalidad de la distribución normal, que consagraron a la curva

normal como una ley de la naturaleza. Galton (1888) escribió:

> Es la suprema ley de la sinrazón. Cada vez que se controla una amplia muestra de elementos caóticos y se les ordena según su magnitud, resultan haber tenido latente todo el tiempo una insospechada y bellísima forma de regularidad.

Seguramente no existe una "ley de la naturaleza" concerniente a la forma que deben asumir las distribuciones de frecuencia. Sin embargo, es cierto que muchas características humanas importantes —tanto físicas como mentales— producen una estrecha aproximación a la curva normal cuando se grafican las medidas de muestras grandes y heterogéneas. Por ejemplo, un hallazgo muy conocido es una curva de distribución casi normal para las características físicas como peso, estatura y tamaño del cerebro al momento del nacimiento (Jensen, 1980).

También se encuentra una distribución aproximadamente normal en el caso de numerosas pruebas mentales, incluso con aquellas que se construyeron totalmente sin referencia a la curva normal. Para ilustrar este punto, se hará referencia a las primeras pruebas diseñadas antes de la fijación psicométrica actual con la distribución normal. Wechsler (1944) eligió los reactivos de la Escala de Inteligencia Wechsler-Bellevue original con base, principalmente, en la variedad de los tipos de reactivo, sin prestar atención a la distribución resultante de las puntuaciones. De hecho, consideraba que la creencia de que las medidas mentales se deben distribuir por sí mismas, según la curva normal, era "equivocada". Sin embargo, cuando graficó la distribución de los CI Totales de su prueba, surgió la previsible distribución casi normal (figura 3–4). Lindvall (1967) encontró lo mismo cuando trazó la gráfica de los datos de la *Pintner Ability Test* (Prueba de Capacidad de Pintner) de 1923. Por tanto, se ve que incluso en ausencia de ajustes psicométricos, la distribución de las puntuaciones de una prueba mental en las muestras de estandarización se aproxima de manera típica a una curva normal.

ASIMETRÍA

La **asimetría** se refiere a la simetría o asimetría de una distribución de frecuencia. Si las puntuaciones de prueba se agrupan hacia el

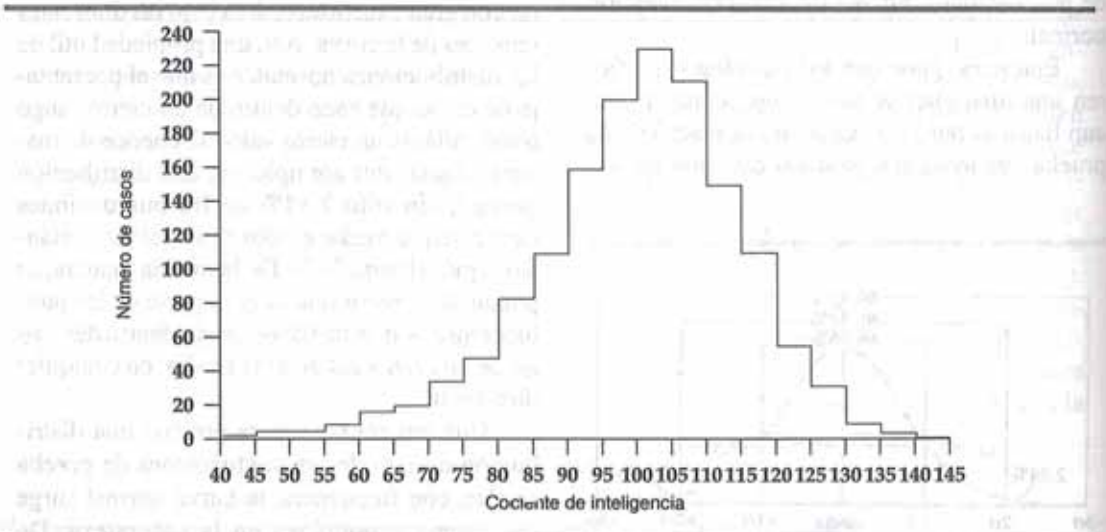
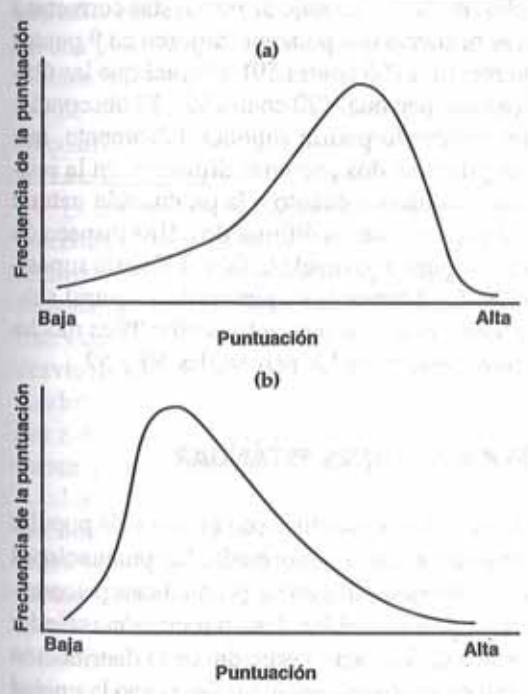


Figura 3–4. Distribución aproximadamente normal de 1 508 CI Totales en la Escala Wechsler-Bellevue. **Fuente:** Reproducido con autorización de Wechsler, D. (1994). *The measurement of adult intelligence* (3a. edición). Baltimore: Williams y Wilkins.

extremo inferior de la escala, se dice que la distribución tiene una asimetría positiva. En el caso opuesto, cuando las puntuaciones de prueba se agrupan en el extremo superior de la escala, se dice que la distribución es asimétrica negativamente (figura 3–5).

En pruebas psicológicas, las distribuciones asimétricas significan, por lo general, que quien desarrolló la prueba ha incluido muy pocos reactivos fáciles o muy pocos reactivos difíciles. Por ejemplo, cuando las puntuaciones en la muestra de estandarización se acumulan en el extremo inferior (asimetría positiva), es probable que la prueba contenga muy pocos reactivos fáciles como para discriminar de manera efectiva en este extremo de la escala. En este caso, los individuos que obtienen puntuaciones de cero o casi cero, en realidad podrían diferir con respecto a la dimensión medida. Sin embargo, la prueba no puede detectar estas diferencias, dado que la mayor parte de los reactivos son demasiado difíciles para estas personas. Por supuesto, también ocurre el patrón contrario. Si las puntuaciones se agrupan al extremo superior (asimetría negativa), es probable que la prueba contenga muy pocos reactivos difíciles como para lograr discriminaciones efectivas en este extremo de la escala.

Cuando la investigación inicial indica que un instrumento produce resultados asimétricos en la muestra de estandarización, por lo común, los autores reforman la prueba a nivel de los reactivos. La solución más directa consiste en añadir reactivos o modificar los existentes, de modo que la prueba tenga más reactivos fáciles (para reducir la asimetría positiva) o más difíciles (para reducir la asimetría negativa). Si es demasiado tarde para revisar el instrumento, el autor de la prueba puede utilizar una transformación estadística para ayudar a producir una distribución más normal de puntuaciones (véase después). Sin embargo, la estrategia preferida consiste en revisar la prueba, de modo que la asimetría sea mínima o inexistente.



a) Asimetría negativa.
b) Asimetría positiva.

Figura 3–5. Curvas Asimétricas de distribución.

TRANSFORMACIÓN DE LAS PUNTUACIONES NATURALES

Darle sentido a los resultados de prueba es, en gran medida, una cuestión de transformar las puntuaciones naturales en formas más interpretables y útiles de información. En el análisis anterior acerca de las distribuciones normales, se insinuaron las transformaciones al mostrar la manera en que el conocimiento sobre la media y la desviación estándar de dichas distribuciones pueden ayudar a determinar la posición relativa de una puntuación individual. En esta sección se continuará con este tema de una manera más directa, al presentar los requisitos formales para varios tipos de transformaciones de las puntuaciones naturales.

PERCENTILES Y RANGOS PERCENTILES

Un **percentil** expresa el porcentaje de personas dentro de la muestra de estandarización que

obtuvieron puntuaciones por debajo de una puntuación natural específica. Por ejemplo, en la prueba de vocabulario representada en el cuadro 3–2, 94% de la muestra se colocó por debajo de la puntuación natural de 25. Así, una puntuación natural de 25 correspondería a un percentil 94, indicado con P_{94}. Obsérvese que los percentiles más altos indican puntuaciones más altas. En el caso extremo, un individuo examinado que haya obtenido una puntuación natural que excediera a todas las puntuaciones en la muestra de estandarización recibiría un percentil de 100 o P_{100}.

Se advierte al lector que no debe confundir los percentiles con el porcentaje de respuestas correctas. Recuérdese que un percentil indica sólo cómo se compara a un individuo con la muestra de estandarización y no transmite el porcentaje de preguntas respondidas de manera correcta. Es posible que, en una prueba difícil, una puntuación natural de 50% de respuestas correctas podría traducirse a un percentil 90, 95 o incluso 100. A la inversa, en una prueba fácil, una puntuación natural de 95% de respuestas correctas podría traducirse a un percentil 5, 10 o 20.

Los percentiles también pueden concebirse como rangos en un grupo de 100 sujetos representativos, donde 1 es el rango inferior y 100 el superior. Obsérvese que los rangos percentiles son el reverso completo de los procedimientos comunes de clasificación por rangos. Un rango percentil (RP) de 1 se encuentra en el extremo inferior de la muestra, mientras que un RP de 99 se encuentra cerca del tope.

Un percentil 50 (P_{50}) corresponde a la mediana o puntuación natural a la mitad de la distribución. Un percentil 25 (P_{25}) con frecuencia se denomina $Q1$ o primer cuartil, debido a que una cuarta parte de las puntuaciones cae por debajo de ese punto. De la misma manera, un percentil 75 (P_{75}) se indica como $Q3$ o tercer cuartil, debido a que tres cuartas partes de las puntuaciones caen por debajo de ese punto.

Los percentiles son fáciles de calcular y son intuitivamente atractivos para los legos y profesionales por igual. Por tanto, no es de sorprender que los percentiles sean el tipo más común de transformación de las puntuaciones naturales que se encuentra en las pruebas psicológicas. Casi cualquier tipo de resultados de prueba se puede presentar como percentiles, incluso cuando otras transformaciones son la meta principal de la prueba. Por ejemplo, las pruebas de inteligencia se utilizan para obtener puntuaciones de CI —un tipo de transformación que se analiza adelante— pero también producen puntuaciones en percentil. Así, un CI de 130 corresponde a un percentil 98, lo cual significa que la puntuación no sólo se encuentra muy por encima del promedio sino, de manera más precisa, que supera a 98% de la muestra de estandarización.

Las puntuaciones percentiles tienen un grave inconveniente: distorsionan la escala de medición subyacente, en especial en los extremos. Un ejemplo específico servirá para aclarar este punto. Considérese un caso hipotético donde cuatro personas obtienen los siguientes percentiles en una prueba: 50, 59, 90 y 99. (Recuérdese que aquí se está hablando de percentiles, no del porcentaje de respuestas correctas.) Las primeras dos personas difieren en 9 puntos percentiles (50 contra 59), al igual que las últimas dos personas (90 contra 99). El observador no entrenado podría suponer, falsamente, que las primeras dos personas difirieron en la misma cantidad en cuanto a la puntuación natural subyacente que las últimas dos. Una inspección de la figura 3–6 revela la falacia de esta suposición. La diferencia en puntuación natural subyacente entre los percentiles 90 y 99 es mucho mayor que entre los percentiles 50 y 59.

PUNTUACIONES ESTÁNDAR

Aunque los percentiles son el tipo más popular de puntuación transformada, las puntuaciones estándar ejemplifican las propiedades psicométricas más deseables. Una puntuación estándar utiliza la desviación estándar de la distribución total de puntuaciones naturales como la unidad fundamental de medida. La **puntuación estándar** expresa la distancia de la media en unidades de desviación estándar. Por ejemplo, una pun-

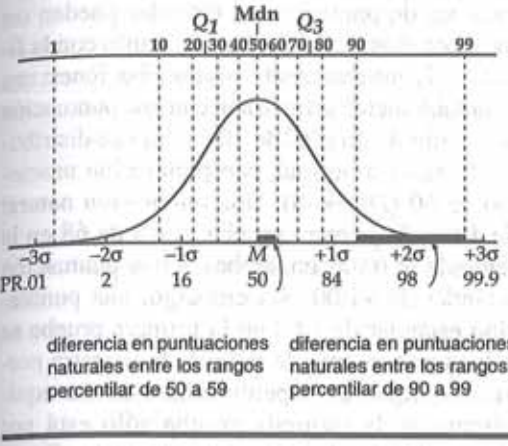


Figura 3–6. Rangos percentiles en una distribución normal.

tuación natural que se encuentra exactamente a una desviación estándar por encima de la media se convierte en una puntuación estándar de +1.00. Una puntuación natural que se encuentra exactamente a la mitad de una desviación estándar por debajo de la media, se convierte en una puntuación estándar de –0.50. Así, una puntuación estándar no sólo expresa la magnitud de la desviación con respecto a la media, sino también la dirección de esa desviación (positiva o negativa).

El cálculo de la puntuación estándar de un individuo (también llamada puntuación *z*) es sencillo: se resta la puntuación natural de la persona examinada a la media del grupo normativo y después se divide esta diferencia entre la desviación estándar del grupo normativo. El cuadro 3–3 ilustra el cálculo de las puntuaciones *z* de tres sujetos con una capacidad sumamente variable en una prueba hipotética.

La puntuación estándar posee la propiedad psicométrica deseable de conservar las magnitudes relativas de distancia entre los valores sucesivos, encontradas entre las puntuaciones naturales originales. Esto se debe a que la distribución de las puntuaciones estándar tiene la misma forma que la distribución de las puntuaciones naturales. Como consecuencia, el uso de las puntuaciones estándar no distorsiona la escala de medida subyacente. Esta fidelidad de la

Cuadro 3–3. Cálculos de las puntuaciones estándar en una prueba hipotética

Para la muestra normativa: $\overline{X} = 50$, $DE = 8$

Puntuación estándar = $z = \dfrac{X - \overline{X}}{DE}$

Persona A: puntuación natural de 35 (debajo del promedio)

$$z = \frac{35 - 50}{8} = -1.88$$

Persona B: puntuación natural de 50 (exactamente promedio)

$$z = \frac{50 - 50}{8} = 0.00$$

Persona C: puntuación natural de 70 (por encima del promedio)

$$z = \frac{70 - 50}{8} = +2.50$$

escala de medición transformada es una de las principales ventajas de las puntuaciones estándar, por encima de los percentiles y de los rangos percentiles. Como se señaló antes, las puntuaciones percentiles distorsionan en gran medida los datos, en especial a los extremos.

Un ejemplo específico servirá para ilustrar este aspecto no distorsionante de las puntuaciones estándar. Considérense cuatro puntuaciones naturales de 55, 60, 70 y 80 en una prueba con una media de 50 y desviación estándar de 10. Las primeras dos puntuaciones difieren en cinco puntos de puntuación natural, mientras las últimas dos difieren en 10 puntos de puntuación natural —dos veces la diferencia del primer par. Cuando las puntuaciones naturales se convierten a puntuaciones estándar, los resultados son +0.50, +1.00, +2.00 y +3.00, respectivamente. El lector observará que las primeras dos puntuaciones difieren en 0.50 desviaciones estándar, mientras las últimas dos difieren en 1.00 desviaciones estándar —dos veces la diferencia del primer par. Así, las

puntuaciones estándar siempre conservan la magnitud relativa de diferencia que se encuentra entre las puntuaciones naturales originales.

Las distribuciones de puntuación estándar poseen importantes propiedades psicométricas que no existen en las distribuciones de puntuación natural. Cuando cada una de estas puntuaciones en una distribución se transforman a una puntuación estándar, el conjunto resultante de puntuaciones estándar siempre tiene una media de cero y una varianza de 1.00. Dado que la desviación estándar es la raíz cuadrada de la varianza, la desviación estándar de las puntuaciones estándar ($\sqrt{1.00}$) es necesariamente 1.00.

Una razón para transformar las puntuaciones naturales en puntuaciones estándar consiste en representar los resultados de pruebas diferentes según una escala común. Si dos distribuciones de puntuaciones de prueba poseen la misma forma, se pueden hacer comparaciones directas de las puntuaciones naturales al transformarlas en puntuaciones estándar. Supóngase, por ejemplo, que un estudiante de primer ingreso a la universidad obtuvo 125 puntos en puntuación natural en una prueba de pensamiento espacial, donde la muestra normativa tuvo un promedio de 100 puntos (con *DE* de 15 puntos). Además, supóngase que obtuvo 110 puntos en puntuación natural en una prueba de vocabulario donde la muestra normativa tuvo un promedio de 90 puntos (con *DE* de 20 puntos). ¿En cuál área de habilidad muestra mayor aptitud, pensamiento espacial o vocabulario?

Si las muestras normativas de ambas pruebas produjeron distribuciones de puntuaciones de la misma forma, se pueden comparar las puntuaciones de pensamiento espacial con las de vocabulario al convertirlas en puntuaciones estándar. La puntuación estándar del estudiante en pensamiento espacial es (125–100)/15 o +1.67, mientras que su puntuación estándar en vocabulario es (110–90)/20 o +1.00. En relación con las muestras normativas, el estudiante tiene mayor aptitud para pensamiento espacial que para vocabulario.

Pero es apropiado hacer una advertencia con respecto a la comparación de puntuaciones estándar de dos distribuciones diferentes. Si éstas no tienen la misma forma, las comparaciones de puntuaciones estándar pueden ser muy engañosas. Se ilustra este punto con la figura 3–7, que representa dos distribuciones: una es notablemente asimétrica, con una puntuación promedio de 30 (*DE* de 10), y otra se distribuye de manera normal, con puntuación promedio de 60 (*DE* de 8). Una puntuación natural de 40 en la primera prueba y una de 68 en la segunda se traducen, ambas, a una puntuación estándar de +1.00. Sin embargo, una puntuación estándar de 1.00 en la primera prueba se coloca por encima de 92% de la muestra normativa, mientras la puntuación estándar equivalente en la segunda prueba sólo está por encima de 84% de la muestra normativa. Cuando dos distribuciones de puntuaciones de prueba no poseen la misma forma, las puntuaciones estándar equivalentes no significan posiciones comparables dentro de las muestras normativas respectivas.

PUNTUACIONES *T* Y OTRAS PUNTUACIONES ESTANDARIZADAS

Muchos psicólogos y educadores aprecian las propiedades psicométricas de las puntuaciones estándar, pero consideran las fracciones decimales y los signos positivos y negativos (p. ej., $z = -2.32$) como distracciones innecesarias. En respuesta a estas preocupaciones, los especialistas en pruebas han diseñado cierto número de variaciones para las puntuaciones estándar que colectivamente se conocen como **puntuaciones estandarizadas**.

Desde un punto de vista conceptual, las puntuaciones estandarizadas son idénticas a las puntuaciones estándar. Ambas contienen exactamente la misma información. No se afecta la forma de la distribución de puntuaciones y la relación entre las puntuaciones estándar y estandarizadas siempre se traza como una línea recta. Sin embargo, las puntuaciones estandarizadas siempre se expresan como números enteros positivos (no existen fracciones decimales ni signos negativos), de modo que muchos usuarios de prueba prefieren representar los resultados en esta forma.

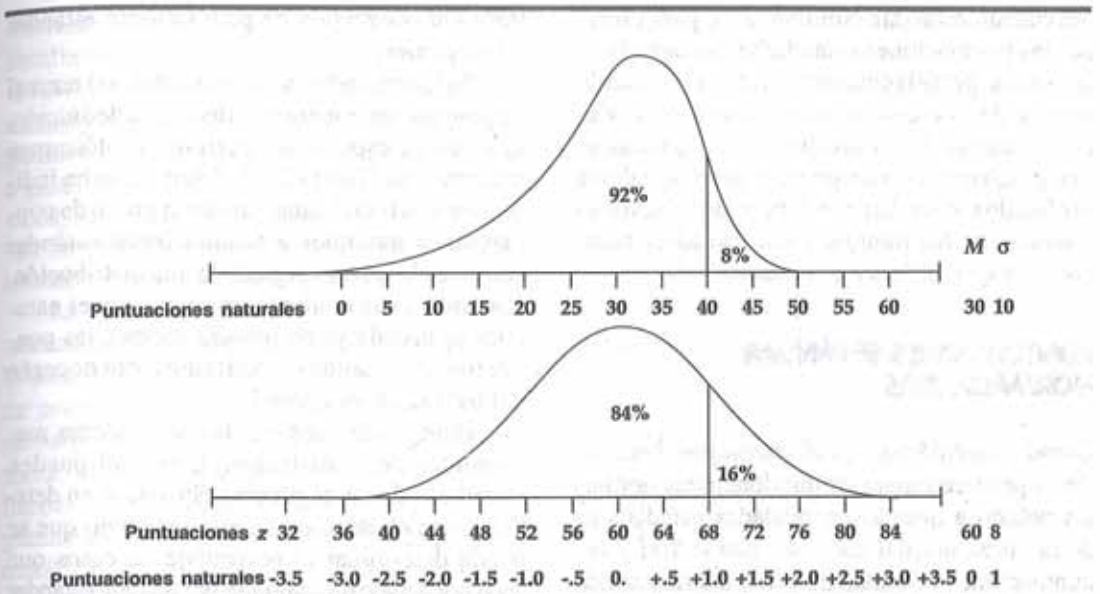


Figura 3–7. Relaciones entre puntuaciones naturales, puntuaciones *z* y posición relativa para dos distribuciones notablemente diferentes.

Las puntuaciones estandarizadas eliminan las fracciones y los signos negativos al producir valores diferentes a cero para la media y 1.00 para la desviación estándar de las puntuaciones transformadas. La media de las puntuaciones transformadas puede establecerse en cualquier valor conveniente como 100 o 500 y la desviación estándar en, digamos, 15 o 100. El punto importante acerca de las puntuaciones estandarizadas es que se puede transformar cualquier distribución a una escala preferida con media y desviación estándar predeterminadas.

Un tipo muy conocido de puntuación estandarizada es la **puntuación *T***, que tiene una media de 50 y desviación estándar de 10. Las escalas en puntuación *T* son especialmente comunes en las pruebas de personalidad. Por ejemplo, en el MMPI, cada escala clínica (como Depresión, Paranoia) se convierte a una medida común, donde 50 es la puntuación promedio y 10 es la desviación estándar de la muestra normativa.

Para transformar las puntuaciones naturales en puntuaciones *T* se utiliza la siguiente fórmula:

$$T = \frac{10\,(X - \bar{X})}{DE} + 50$$

El término $(X-\bar{X})/DE$ es, por supuesto, equivalente a *z*, de modo que se puede volver a escribir la ecuación para *T* como una simple transformación de *z*:

$$T = 10z + 50$$

Para cualquier distribución de puntuaciones naturales, las puntuaciones *T* correspondientes tendrán un promedio de 50. Además, para la generalidad de las distribuciones, gran parte de las puntuaciones *T* caerán entre los valores de 20 y 80; es decir, dentro de tres desviaciones estándar a partir de la media. Por supuesto, es totalmente posible que existan puntuaciones *T* fuera de este rango y, quizá, incluso son probables en poblaciones especiales. En entornos clínicos, no es poco común observar puntuaciones *T* muy altas —incluso tan altas como 90— en inventarios de personalidad como el MMPI.

Las puntuaciones estandarizadas se pueden adaptar para producir cualquier media y

desviación estándar. Sin embargo, para eliminar las puntuaciones estandarizadas negativas, la media preseleccionada debe ser cuando menos cinco veces tan grande como la desviación estándar. En la práctica, quienes desarrollan pruebas dependen de unos cuantos valores preferidos para las medias y desviaciones estándar de las puntuaciones estandarizadas, como se ejemplifica en el cuadro 3–4.

PUNTUACIONES ESTÁNDAR NORMALIZADAS

Como se señaló antes, los psicólogos y educadores prefieren manejar distribuciones normales debido a que las propiedades estadísticas de la curva normal son muy conocidas, y las puntuaciones estándar de estas distribuciones se pueden comparar directamente. Quizás el lector se ha preguntado qué recurso se tiene disponible para el autor de una prueba que encuentra que ésta produce una distribución asimétrica de puntuaciones en la muestra normativa. Por fortuna, las distribuciones de puntuaciones asimétricas o que no son normales en otros sentidos pueden transformarse o normalizarse para encajar con una curva normal. Aunque los especialistas en pruebas han diseñado varios métodos para convertir una distribución no normal en normal, se analizará sólo el enfoque más popular —la conversión de percentiles a puntuaciones estándar normalizadas. Por extraño que parezca, es más fácil explicar este enfoque si primero se describe el proceso contrario, la conversión de puntuaciones estándar a percentiles.

Se ha señalado que una distribución normal de puntuaciones naturales tiene, por definición, una forma específica, definida en términos matemáticos (figura 3–3). Además, se ha indicado que la transformación de un grupo de puntuaciones naturales a puntuaciones estándar conserva la forma original de una distribución. Por tanto, si un conjunto de puntuaciones naturales se distribuye de manera normal, las puntuaciones estándar resultantes obedecerán también la curva normal.

También se sabe que las propiedades matemáticas de la distribución normal pueden calcularse de modo preciso. Sin entrar en detalles sobre el cálculo, debería ser obvio que se puede determinar el porcentaje de casos que caen por debajo de cualquier puntuación estándar particular. Por ejemplo, en la figura 3–6, una puntuación estándar de –2.00 (designada como -2σ) se encuentra por encima de 2.14% de los casos. Así, una puntuación estándar de –2.00 corresponde al percentil 2.14. De la misma forma, cualquier puntuación estándar posible puede expresarse en términos de su percentil correspondiente. El Apéndice D enumera los percentiles para las puntuaciones estándar y para varias otras puntuaciones transformadas.

La obtención de **puntuaciones estándar normalizadas** se logra mediante el cálculo inverso. A saber, se utiliza el percentil de cada puntuación natural para determinar su puntuación estándar correspondiente. Si se hace esto con todos y cada uno de los casos en una distribución

Cuadro 3–4. Medias y desviaciones estándar de puntuaciones estandarizadas comunes

Tipo de medida	Ejemplos específicos	Media	Desviación estándar
CI Total	WAIS-III	100	15
Subescalas de prueba de CI	Vocabulario, Diseño con Cubos	10	3
Escalas de Prueba de Personalidad	Depresión, Paranoia en el MMPI-2	50	10
Pruebas de aptitud	Examen de Selección para el Posgrado, Pruebas de Evaluación Escolar	500	100

diferente a la normal, las puntuaciones estándar resultantes se distribuirán en forma normal. Obsérvese que en dicha distribución normalizada de las puntuaciones estándar, dichas puntuaciones no se obtienen de manera directa a partir de la fórmula normal para el cálculo, sino que se determinan de modo indirecto a través de calcular primero el percentil y después obtener la puntuación estándar equivalente.

La conversión de percentiles a puntuaciones estándar normalizadas podría parecer como la solución ideal para el problema de los datos de prueba que no se ajustan a la norma. Sin embargo, existe un inconveniente potencialmente serio: las puntuaciones estándar normalizadas son una transformación no lineal de las puntuaciones naturales. Así, las relaciones matemáticas establecidas con las puntuaciones naturales pueden no resultar ciertas para las puntuaciones estándar normalizadas. En una distribución notablemente asimétrica, incluso podría concebirse que una puntuación natural que se encuentre significativamente por debajo de la media tenga una puntuación estándar normalizada que supere la media.

En la práctica, las puntuaciones estándar normalizadas se utilizan pocas veces. Dichas transformaciones son apropiadas sólo cuando la muestra normativa es grande y representativa y la distribución de puntuaciones naturales se aleja sólo ligeramente de la curva normal. A propósito, la causa más probable para estas distribuciones no normales de las puntuaciones es el nivel inapropiado de dificultad de los reactivos de prueba; por ejemplo, demasiados reactivos difíciles o fáciles.

En este caso nos enfrentamos a una trampa sin salida, en cuanto a que las distribuciones que se alejan ligeramente de la curva normal no cambian en mucho cuando se les normaliza y, por tanto, se obtiene poco en el proceso. De manera irónica, las puntuaciones estándar normalizadas producen el mayor cambio cuando se trata de distribuciones que se alejan mucho de la curva normal. Sin embargo, cuando la distribución de puntuaciones naturales es de este tipo, quienes desarrollan pruebas harían mejor en volver a la mesa de trabajo y ajustar el nivel de dificultad de los reactivos de prueba, de modo que produzcan una distribución normal, en lugar de sucumbir al arreglo estadístico parcial de las puntuaciones estándar normalizadas.

ESTANINAS, ESTENES Y ESCALA C

Por último, se hace una breve mención de tres transformaciones de puntuaciones naturales que tienen principalmente un interés histórico. La Fuerza Aérea de EUA desarrolló la escala de estaninas (estándar nueve) durante la segunda Guerra Mundial. En una **escala de estaninas**, todas las puntuaciones naturales se convierten a un sistema de un solo dígito de puntuaciones que van de 1 a 9. La media de las puntuaciones de estaninas es siempre 5 y la desviación estándar es aproximadamente 2. La transformación de puntuaciones naturales a estaninas es simple: las puntuaciones se clasifican de menor a mayor y 4% más bajo de las puntuaciones se convierte en una estanina de 1, el siguiente 7% se convierte a una estanina de 2 y así sucesivamente (véase cuadro 3–5). La principal ventaja de las estaninas es que se restringen a números de un solo dígito. Esta era una cualidad considerable en la época anterior a las computadoras modernas, cuando los datos se perforaban en tarjetas Hollerith que tenían que llevarse y almacenarse físicamente en repisas. Dado que una estanina se podía perforar en una sola columna, se requerían muchas menos tarjetas que si se ingresaran las puntuaciones naturales originales.

Cuadro 3–5. Porcentajes de la distribución para utilizarse en la conversión a estaninas

Porcentaje	4	7	12	17	20	17	12	7	4
Estaninas	1	2	3	4	5	6	7	8	9

Los expertos en estadística han propuesto diversas variaciones para el tema de la escala de estaninas. Canfield (1951) propuso la **escala estenes** de 10 unidades, con 5 unidades por arriba y 5 por debajo de la media. Guilford y Fruchter (1978) propusieron la **escala C**, que consiste en 11 unidades. Aunque las estaninas aún tienen un amplio uso, variaciones como la escala estenes y la escala C nunca lograron gran interés por parte de los autores de pruebas.

RESUMEN DE LAS NORMAS BASADAS EN ESTADÍSTICAS

Se ha aludido en varias ocasiones la facilidad con la que las puntuaciones estándar, puntuaciones *T*, estaninas y percentiles se pueden transformar unas en otras, en especial si la distribución subyacente de puntuaciones naturales se adapta a la curva normal. De hecho, la forma exacta en la que se presentan las puntuaciones es, en gran medida, una cuestión de convención y preferencia personal. Por ejemplo, un CI de 115 en el WAIS-III puede presentarse también como una puntuación estándar de +1.00 o una puntuación *T* de 60 o un rango percentil 84. Todos estos resultados transmiten exactamente la misma información.* La figura 3–8 resume las

* Un CI de 115 en el WAIS-III también se puede expresar como una estanina de 7 puntos. Sin embargo, es necesario mencionar que se pierden datos cuando se informa en estaninas. Obsérvese que los CI en el rango de 111 a 119 se transforman **todos** en una estanina de 7 puntos. Por tanto, si se dice sólo que un individuo ha logrado una estanina de 7 puntos en una prueba de inteligencia, no se sabe el equivalente exacto en CI.

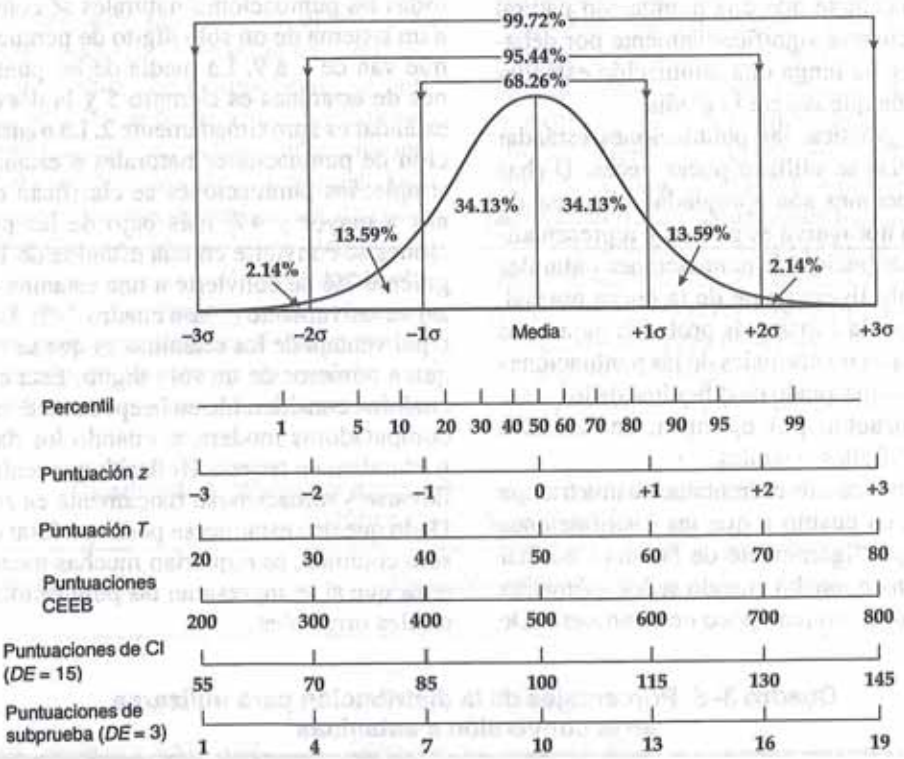


Figura 3–8. Equivalencias entre transformaciones comunes de las puntuaciones naturales en una distribución normal.

relaciones que existen entre las normas con base estadística que se emplean de modo más común.

Esto termina la breve introducción a las muchas técnicas con las cuales se pueden resumir y transformar de manera estadística los datos de prueba de una muestra normativa. Nunca se debe perder de vista el principal propósito de estas transmutaciones estadísticas, es decir, el ayudar al usuario de prueba a darle significado a una puntuación individual en relación con un grupo apropiado de comparación.

Pero, ¿qué es un grupo apropiado de comparación? ¿Qué características se deben requerir en las personas dentro del grupo normativo? ¿Cómo se debe elegir a estos individuos? ¿Cuántas personas se necesitan? Éstas son preguntas importantes que influyen sobre la pertinencia de los resultados de una prueba, de la misma manera que la selección adecuada de los reactivos y los procedimientos estandarizados de prueba. En lo que resta de este tema, se analizarán los procedimientos implicados en la selección de un grupo normativo.

SELECCIÓN DE UN GRUPO NORMATIVO

Cuando se elige un grupo normativo, quienes desarrollan pruebas se esfuerzan por obtener una muestra representativa de la población para la cual se diseñó la prueba (Petersen, Kolen y Hoover, 1989). En teoría, obtener un grupo normativo representativo es simple y sencillo. Considérese una prueba de rendimiento escolar diseñada para niños de sexto grado de primaria en EUA. La población relacionada la constituyen todos los niños en sexto grado, de costa a costa del país y en Alaska y Hawai. Se puede obtener una muestra representativa de estos individuos potenciales a través de un muestreo aleatorio por computadora de aproximadamente 10 000 niños, entre los millones de niños elegibles. Cada uno tendría igual probabilidad de ser elegido para realizar la prueba; es decir, la estrategia de selección sería un simple **muestreo aleatorio**. Los resultados de dicho muestreo constituirían una fuente ideal para los datos normativos. Con una muestra aleatoria amplia, es casi seguro que la diversidad de antecedentes étnicos, clases sociales, localizaciones geográficas y entornos urbanos contra rurales tendrían una representación proporcional en la muestra.

En el mundo real, obtener muestras normativas nunca es tan sencillo y definitivo como el caso hipotético anterior. Los investigadores no tienen una lista completa de todos los niños de sexto grado en el país y, aunque la tuvieran, quienes desarrollan pruebas no podrían obligar a todos los niños seleccionados al azar a que participaran en la estandarización de una prueba. También surgen asuntos relacionados con el costo. Debe pagarse a los psicómetras para que apliquen la prueba al grupo normativo. Los autores de la prueba pueden optar por unos cuantos cientos de personas representativas en vez de utilizar un número mayor.

Para ayudar a asegurar que los grupos normativos de menor tamaño sean verdaderamente representativos de la población para la cual se diseñó la prueba, los autores de la misma utilizan un **muestreo aleatorio estratificado**. Este enfoque consiste en estratificar, o clasificar, a la población blanco en cuanto a variables antecedentes (como edad, sexo, raza, clase social, nivel educativo) y después seleccionar al azar un porcentaje apropiado de personas dentro de cada estrato. Por ejemplo, si 12% de la población relacionada es afroestadounidense, entonces quien desarrolla la prueba elige los sujetos de manera aleatoria, pero con la restricción de que 12% del grupo normativo debe ser también afroestadounidense.

En la práctica, muy pocos creadores de pruebas reproducen completamente el muestreo aleatorio o el muestreo aleatorio estratificado en el proceso de selección del grupo normativo. Lo que es más típico es un esfuerzo de buena fe por elegir una muestra diversa y representativa de escuelas fuertes y débiles; vecindarios minoritarios y blancos; ciudades grandes y pequeñas y comunidades del norte, este, centro y sur. Si esta muestra engloba entonces

aproximadamente el mismo porcentaje de minorías, habitantes de las ciudades, familias de clase alta y baja que el censo nacional, entonces los autores de la prueba se sienten seguros de que el grupo es representativo.

Existe una importante lección en las incertidumbres, concesiones y aspectos prácticos de la selección del grupo normativo; a saber, que las normas de pruebas psicológicas no son absolutas, universales o intemporales. Son relativas a una época histórica y a la población normativa particular de la cual se derivaron. Se ilustrará la naturaleza efímera de las estadísticas normativas en una sección posterior cuando se muestre cómo una importante prueba de CI en cuyas normas se estableció un promedio nacional de 100 puntos en 1974, produjo un promedio nacional de 107 en 1988. Incluso las normas que se seleccionan con gran cuidado y se basan en muestras grandes pueden volverse obsoletas en el curso de un decenio —y en ocasiones menos.

NORMAS POR EDAD Y GRADO ESCOLAR

A medida que uno crece, se cambia de manera calculable, ya sea para mejorar o empeorar. Esto es obviamente cierto en la infancia, cuando las habilidades intelectuales mejoran visiblemente de un mes a otro. En la edad adulta, el cambio personal es lento, pero aún se puede discernir. Por ejemplo, se espera que los adultos muestren un nivel más maduro de vocabulario con cada decenio que transcurre (Gregory y Gernert, 1990).

Una **norma por edad** representa el nivel de desempeño en la prueba para cada grupo independiente según la edad dentro de la muestra normativa. El propósito de las normas por edad es el de facilitar las comparaciones entre individuos de la misma edad. Con las normas por edad, el desempeño de una persona se interpreta en relación con los sujetos dentro de la estandarización que tienen la misma edad. El rango de edades para un grupo normativo por edad puede variar de un mes a un decenio o más, según el grado en que el desempeño en la prueba depende de los años de vida. Para características que cambian con rapidez —como las capacidades intelectuales durante la infancia— los creadores de pruebas podrían informar normas independientes de prueba para grupos de edad definidos de manera estrecha; por ejemplo, intervalos de cuatro meses. Esto permite que, por decir, el examinador compare los resultados de prueba de un niño que tiene cinco años, dos meses de edad (edad 5-2) con la muestra normativa de niños que van de los 5-0 a los 5-4 años. En contraste, las características adultas cambian de manera más lenta y podría ser suficiente con informar datos normativos para intervalos de edad de 5 o 10 años.

Las normas según el grado escolar son conceptualmente similares a las normas por edad. Una **norma por grado** representa el nivel de desempeño en pruebas para cada grado escolar dentro de la muestra normativa. Las normas por grado se utilizan poco en el caso de pruebas de capacidad. Sin embargo, éstas son especialmente útiles en entornos educativos, cuando se presentan los niveles de rendimiento de los niños en edad escolar. Dado que el rendimiento académico en muchas áreas de contenido depende en gran medida de la exposición al currículo de materias según el grado escolar, la comparación de un estudiante contra una muestra normativa del mismo grado es más apropiada que emplear comparaciones basadas en la edad.

NORMAS LOCALES Y DE SUBGRUPO

En el caso de muchas aplicaciones, las normas locales o de subgrupo son necesarias para ajustarse al propósito específico de una prueba. Las **normas locales** se obtienen de individuos representativos a nivel local, por el contrario de una muestra nacional. De la misma manera, las **normas de subgrupo** consisten de las puntuaciones obtenidas por un subgrupo representativo (afroestadounidenses, individuos de origen latinoamericano, mujeres), por el contrario de una muestra diversificada a nivel nacional. Como ejemplo del uso de las normas locales,

el encargado de admisiones de una escuela semisuperior que atrae principalmente a residentes locales podría preferir la consulta de normas estatales en una prueba de aprovechamiento escolar, en vez de las normas nacionales.

Como regla general, cuando un subgrupo identificable tiene un desempeño en prueba que es superior o inferior de manera apreciable al de la muestra de estandarización definida de modo más amplio, puede ser útil construir normas complementarias para ese subgrupo. Los subgrupos pueden formarse con respecto al sexo, antecedentes étnicos, región geográfica, ambiente urbano contra rural, nivel socioeconómico y muchos otros factores.

El que las normas locales o de subgrupo sean benéficas depende del propósito de la prueba. Por ejemplo, las normas étnicas para pruebas estandarizadas de inteligencia pueden ser superiores a las normas con base nacional en cuanto a pronosticar la competencia dentro del ambiente no escolar del niño. Sin embargo, puede ser que las normas étnicas no pronostiquen el éxito que tendrá el niño en los programas educativos de las escuelas públicas convencionales (Mercer y Lewis, 1978). Por tanto, las normas locales y de subgrupo deben utilizarse de manera cautelosa.

CUADROS DE EXPECTATIVAS

Una forma práctica que pueden asumir las normas es un **cuadro de expectativas**, que representa las relaciones establecidas entre las puntuaciones de prueba y el resultado esperado en una tarea relacionada (Harmon, 1989). Los cuadros de expectativas son especialmente útiles con pruebas predictivas utilizadas para pronosticar criterios bien definidos; por ejemplo, podría describir la relación entre las puntuaciones de una prueba de aptitud escolar (medio de predicción) y el subsecuente promedio de calificaciones en la universidad (criterio).

Los cuadros de expectativas se basan siempre en los resultados previos del medio de predicción y del criterio en muestras amplias de personas examinadas. El valor práctico de tabular de esta manera la información normativa es que los nuevos individuos que se examinen recibirán un avance de sus probabilidades de éxito en cuanto al criterio. Por ejemplo, a los estudiantes de educación media superior que realizan una prueba de aptitud escolar, se les pueden presentar las probabilidades estadísticas de lograr un promedio específico de calificaciones en la universidad.

Con base en 7 835 individuos examinados previamente que de manera subsecuente asistieron a una universidad importante, el cuadro de expectativas del cuadro 3–6 proporciona la probabilidad de lograr ciertas calificaciones en el primer año de universidad, en función de las puntuaciones en el examen del *American College Testing* (ACT). La prueba ACT se aplica por lo común a estudiantes del último año de preparatoria que han expresado un interés por asistir a la universidad. La primera columna del cuadro muestra las puntuaciones en la prueba ACT, divididas en 10 intervalos de clase. La segunda columna proporciona el número de estudiantes cuyas puntuaciones caen dentro de cada intervalo. Las ecuaciones restantes en cada fila muestran el porcentaje de estudiantes dentro de cada intervalo de puntuaciones de prueba que de manera subsecuente recibieron calificaciones universitarias dentro de un rango designado. De los 117 estudiantes que obtuvieron de 31 a 33 puntos en la ACT, sólo 2% recibió en su primer año de universidad un promedio de calificaciones por debajo de 1.50, mientras que 64% obtuvo calificaciones superlativas de 3.50 a una "A" perfecta o 4.00.* En el otro extremo, de los 102 estudiantes que tuvieron puntuaciones inferiores a 10 puntos en la ACT, 80% completo (60% más 20%), recibió calificaciones universitarias durante el primer año que se encontraban por debajo de un promedio de 2.00 o "C".

*Nota del traductor: Las calificaciones universitarias en EUA abarcan sólo cuatro puntos. Es decir, la puntuación de 1.5 equivaldría a una calificación muy baja; un 3.5 a una alta y 4.00 es la puntuación más alta posible. Estas calificaciones se representan también con letras, con las equivalencias aproximadas que menciona el autor.

Cuadro 3–6. Cuadro de expectativas que muestra la relación entre la puntuación mixta de la prueba ACT y las calificaciones del primer año de universidad para 7 835 alumnos de una universidad estatal importante

Puntuaciones en la prueba ACT	Número de casos	*Promedio de Calificaciones (Escala 4.00)* 0.00–1.49	1.50–1.99	2.00–2.49	2.50–2.99	3.00–3.49	3.50–4.00
34–36	3	0	0	33	0	0	67
31–33	117	2	2	4	9	19	64
28–30	646	10	6	10	17	23	35
25–27	1 458	12	10	16	19	24	19
22–24	1 676	17	10	22	20	20	11
19–21	1 638	23	14	25	18	16	4
16–18	1 173	31	17	24	15	11	3
13–15	690	38	18	25	12	6	1
10–12	332	54	16	20	6	3	1
inferior a 10	102	60	20	13	8	0	0

Nota: algunas filas totalizan más de 100% debido a errores de redondeo.

Fuente: Cortesía de Archie George, Management Information Services, Universidad de Idaho.

Por supuesto, los cuadros de expectativa no predeterminan el éxito o fracaso de las personas examinadas con respecto al criterio. En un caso individual, es posible que un estudiante con una baja calificación en la ACT pueda vencer el pronóstico y obtener un promedio de calificaciones de 4.00 en la universidad. Sin embargo, es más común que los nuevos individuos examinados descubran que los cuadros de expectativa proporcionan un anticipo bastante preciso del desempeño en el criterio.

Pero existen algunos casos individuales en los que los cuadros de expectativas pueden volverse imprecisos. Un cuadro de expectativas se basa siempre en el desempeño previo de una muestra grande y representativa de personas evaluadas, cuya ejecución en prueba y resultados en el criterio reflejaron condiciones sociales y políticas institucionales existentes. Si hay un cambio en políticas o condiciones, un cuadro de expectativas puede volverse obsoleto y engañoso. Considérese el cuadro de expectativas en la figura 3–9, que presenta la probabilidad de terminar la secundaria en función del CI en el séptimo grado escolar (Dillon, 1949, citado en Matarazzo, 1972, p. 283). Obsérvese que en el decenio de 1940, sólo 4% de los estudiantes de séptimo grado con un CI por debajo de 85 terminaron la secundaria. Sin embargo, las políticas sociales y ambientes escolares han cambiado desde ese entonces. En la actualidad se da un fuerte énfasis a los servicios sociales dirigidos a los estudiantes discapacitados, con la meta de retenerlos y que lleguen a graduarse. Como resultado, el cuadro de expectativas de la figura 3–9 seguramente sería erróneo y pesimista si se aplicara a los estudiantes actuales de séptimo grado que tienen un CI bajo.

PRUEBAS REFERIDAS A CRITERIO

Esta unidad termina con una breve mención de una alternativa a las pruebas referidas a la norma; es decir, las pruebas referidas a criterio (Frechtling, 1989; Glaser, 1963; Popham, 1978; Berk, 1984). Mientras que las primeras utilizan

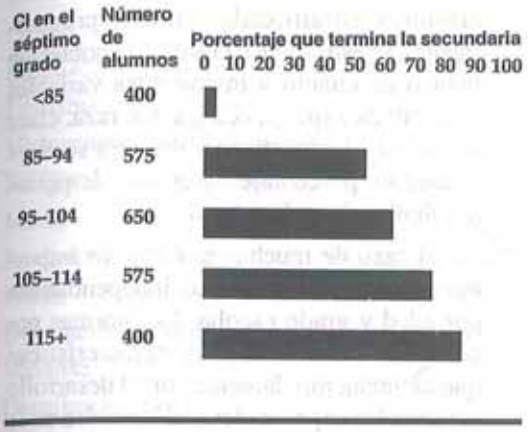


Figura 3–9. Expectativa de graduación de la secundaria en función del CI en el séptimo grado escolar. **Fuente:** Basado en datos de Dillon, H. J. (1949). *Early school leavers: A major educational problem.* Nueva York: *National Child Labor Committee.* Citado en Matarazzo (1972).

una muestra representativa de personas como su marco de referencia interpretativo, una prueba referida a criterio compara los logros de un individuo examinado con un área bien definida de contenido. Por ejemplo, el desempeño de la persona que se somete a prueba podría presentarse como un nivel específico de habilidad en aritmética (suma correctamente 2 números de 3 dígitos en 100% de las ocasiones, pero sólo tiene 70% de precisión cuando suma 3 números de 3 dígitos).

Las pruebas referidas a criterio representan un cambio fundamental en perspectiva. El centro de atención se coloca en aquello que el examinado puede hacer, más que en comparaciones con los niveles de desempeño de otros individuos. Así, las pruebas referidas a criterio identifican el dominio (o falta de dominio) absoluto de la persona examinada en relación con conductas específicas. En general, estas pruebas se han utilizado en entornos escolares para ayudar a los maestros dentro del salón de clase a determinar qué se debe enseñar y también para evaluar los efectos precisos de los esfuerzos de enseñanza del currículo escolar.

Las pruebas referidas a criterio son más adecuadas para examinar las habilidades académicas básicas (p. ej., nivel de lectura, habilidades de cálculo aritmético) en ambientes educativos. Sin embargo, estos tipos de instrumentos son sumamente inapropiados para examinar capacidades de nivel superior, debido a que es difícil formular objetivos específicos para dichas áreas de contenido. Considérese un caso en particular: ¿cómo se desarrollaría una prueba referida a criterio para examinar la pericia en programación de computadoras? Sería difícil proponer conductas específicas que pudieran poseer todos los expertos en programación de computadoras y, por tanto, sería casi imposible construir una prueba referida a criterio para esta habilidad de alto nivel. Berk (1984) analiza los problemas técnicos en la construcción y evaluación de este tipo de pruebas.

RESUMEN

1. Un grupo normativo consiste de una muestra de individuos examinados que son representativos de la población hacia la cual se dirige la prueba. Una distribución de frecuencias es útil para representar la distribución de las puntuaciones de prueba dentro de ciertos intervalos de puntuación para un grupo normativo. Un histograma es la representación gráfica de una distribución de frecuencia.
2. Las medidas de tendencia central para conjuntos de puntuaciones incluyen la media o promedio aritmético, la mediana o calificación a la mitad de las puntuaciones clasificadas, y la moda, que es la puntuación que ocurre con mayor frecuencia.
3. Las medidas de variabilidad para un grupo de puntuaciones incluyen la varianza y su raíz cuadrada, la desviación estándar, que es la medida preferida en las pruebas psicológicas.

Estos índices ayudan a estimar la dispersión de las puntuaciones a través de incorporar en sus fórmulas las sumas de las desviaciones con respecto a la puntuación media, elevadas al cuadrado.

4. La distribución de puntuaciones de prueba de grandes grupos de individuos heterogéneos se asemeja con frecuencia a la distribución normal; una curva simétrica, definida en términos matemáticos y que adopta forma de campana. Los psicólogos prefieren tratar con puntuaciones de prueba que se distribuyen normalmente, debido a que las características estadísticas de la distribución normal son muy conocidas.
5. Una distribución asimétrica es aquélla en la que las puntuaciones se agrupan en el extremo inferior (asimetría positiva) o en el extremo superior (asimetría negativa). En las pruebas psicológicas, la causa más común de asimetría positiva es la presencia de muy pocos reactivos fáciles, mientras la causa más común de asimetría negativa es que la prueba tiene muy pocos reactivos difíciles.
6. Un percentil expresa el porcentaje de personas dentro de la muestra de estandarización que tuvieron puntuaciones por debajo de cierta puntuación natural. Los percentiles varían de 0 a 100. Es importante distinguir entre el percentil (una medida relativa) y el porcentaje de respuestas correctas (una medida absoluta).
7. Una puntuación estándar expresa la puntuación natural de la persona examinada en términos de su distancia con respecto a la media en unidades de desviación estándar. La fórmula para una puntuación estándar es $z = (X - \overline{X})/DE$. La puntuación T es una puntuación estándar con media de 50 y desviación estándar de 10. La fórmula para una puntuación T es:

$$T = 10\,(X - \overline{X})/DE + 50$$

8. El enfoque más común para seleccionar un grupo normativo es a través del muestreo aleatorio estratificado. En este procedimiento, se estratifica o clasifica la población blanco en cuanto a importantes variables antecedentes (p. ej., edad, sexo, raza, clase social, nivel educativo) y después se elige al azar un porcentaje apropiado de personas dentro de cada estrato.
9. En el caso de muchas pruebas, es importante proporcionar normas independientes por edad y grado escolar. Las normas por edad son necesarias para características que cambian rápidamente con el desarrollo —como las capacidades intelectuales en la infancia. Las normas por grado se utilizan comúnmente en entornos escolares cuando se informa sobre los niveles de rendimiento de niños en edad escolar.
10. Las normas locales y de subgrupo pueden ser valiosas si un subgrupo identificable tiene un desempeño en una prueba que, de manera evidente, es mejor o peor que el de la muestra de estandarización definida en términos más amplios.
11. Un cuadro de expectativas —una forma de estandarización de prueba— representa la relación establecida entre las puntuaciones de prueba y el resultado esperado en una tarea relacionada. Por ejemplo, un cuadro de expectativas podría mostrar la relación entre las puntuaciones de una prueba de aptitud escolar y el subsecuente promedio de calificaciones en la universidad.
12. Una prueba referida a criterio compara los logros del individuo en la prueba con un área bien definida de contenido. Estas pruebas ayudan a identificar el dominio o falta de dominio de la persona examinada en cuanto a conductas específicas. Por ejemplo, los resultados de una prueba referida a criterio podrían especificar que la persona suma correctamente 2 números de 3 dígitos en 100% de las ocasiones.

TÉRMINOS Y CONCEPTOS CLAVE

Grupo normativo	p. 70	Puntuación *T*	p. 81
Puntuación natural	p. 70	Puntuaciones estandararizadas normalizadas	p. 82
Distribución de frecuencias	p. 72	Escala de estaninas	p. 83
Histograma	p. 73	Escala estenes	p. 84
Polígono de frecuencias	p. 73	Escala C	p. 84
Media	p. 73	Muestreo aleatorio	p. 85
Mediana	p. 73	Muestreo aleatorio estratificado	p. 85–86
Moda	p. 73	Norma por edad	p. 86
Desviación estándar	p. 74	Norma por grado	p. 86
Varianza	p. 74	Normas locales	p. 86
Distribución normal	p. 74–75	Normas de subgrupo	p. 86–87
Asimetría	p. 76–77	Cuadro de expectativas	p. 87
Percentil	p. 77–78		
Puntuaciones estandarizadas	p. 80		

TEMA 3B
Conceptos de confiabilidad

Ejemplo de caso 3–2. Confiabilidad de prueba y testimonio en un tribunal
Teoría clásica y fuentes de error de medición
Fuentes de error de medición
Error de medición y confiabilidad
Coeficiente de confiabilidad
Coeficiente de correlación
Coeficiente de correlación como coeficiente de confiabilidad
Confiabilidad como estabilidad temporal
Confiabilidad como consistencia interna
Modelo de muestreo del dominio y la teoría de la capacidad para la generalización
Circunstancias especiales en la estimación de la confiabilidad
Interpretación de los coeficientes de confiabilidad
Confiabilidad y error estándar de medida
Resumen
Términos y conceptos clave

La confiabilidad se refiere al atributo de consistencia en medición. Sin embargo, es poco frecuente que la confiabilidad sea un asunto de todo o nada; lo más común es que sea una cuestión de grado. Muy pocas medidas de las características físicas o psicológicas son totalmente consistentes, incluso de un momento al siguiente. Por ejemplo, una persona que se pesa en una báscula dos veces, en rápida sucesión, podría registrar un peso de 66 kg la primera vez y 66 kg, 100 g, la segunda. El mismo individuo podría responder dos formas supuestamente equivalentes de una prueba de CI y obtener 114 puntos en la primera y 119 en la segunda. Dos medidas sucesivas de velocidad de respuesta —oprimir con rapidez una tecla cada vez que aparece la letra X en una pantalla de microcomputadora— podrían producir un tiempo de reacción de 223 milisegundos en el primer ensayo y de 341 milisegundos en el siguiente. En estos ejemplos se observa un patrón de consistencia —los pares de mediciones no son totalmente aleatorios— pero también es evidente que existen cantidades diferentes de inconsistencia. En el corto plazo, las medidas de peso son muy consistentes, las puntuaciones de pruebas intelectuales son moderadamente estables, pero el tiempo de reacción simple es un tanto errático.

Es mejor considerar al concepto de **confiabilidad** como un continuo que abarca desde la consistencia mínima de una medición (p. ej., tiempo de reacción simple) a la casi perfecta repetibilidad de los resultados (p. ej., peso). La mayor parte de las pruebas psicológicas se encuentran en algún sitio entre estos dos extremos. Con respecto a las pruebas, un grado aceptable de confiabilidad es más que una cuestión académica. Después de todo, sería absurdo y poco ético basar las decisiones importantes en resultados de prueba que no pueden repetirse (ejemplo de caso 3–2).

Los psicómetras han diseñado varios métodos estadísticos para estimar el grado de confiabilidad de las mediciones y se explorará con cierto detalle el cálculo de dichos coeficientes de confiabilidad. Pero, en primer término, se analiza un problema más fundamental con el propósito de ayudar a aclarar el significado de la confiabilidad: ¿cuáles son las fuentes de consistencia e inconsistencia en los resultados de una prueba psicológica?

TEORÍA CLÁSICA Y FUENTES DE ERROR DE MEDICIÓN

La teoría de la medición presentada aquí se ha llamado teoría clásica, debido a que se le elaboró a partir de suposiciones simples realizadas por los teóricos de las pruebas desde el origen de estos instrumentos. Este enfoque también se denomina **teoría de puntuaciones verdaderas y falsas**, por razones que se explican más adelante. Charles Spearman (1904) estableció las bases para la teoría, que después aumentaron y revisaron algunos psicólogos contemporáneos (Feldt y Brennan, 1989; Gulliksen, 1950; Lord y Novick, 1968; Kline, 1986). Se debe mencionar que no existe un modelo rival. La teoría del muestreo del dominio surgió con la reformulación de Tryon (1957), que Cronbach, Gleser, Nanda y Rajaratnam (1972) han extendido recientemente para formar una teoría alternativa denominada teoría de la capacidad para la generalización. Más adelante, se hará una breve referencia a la perspectiva del muestreo del dominio. Sin embargo, la teoría clásica aún es el punto de vista prevaleciente en psicometría. En vista de ello, nos concentraremos en este modelo.

El punto básico inicial de la **teoría clásica de medición** es la idea de que las puntuaciones de prueba son el resultado de la influencia de dos factores:

1. Factores que contribuyen a la consistencia. Éstos comprenden en su totalidad los atributos estables del individuo que el examinador intenta medir.
2. Factores que contribuyen a la inconsistencia. Éstos incluyen características del individuo, la prueba o la situación que no tienen nada que ver con el atributo a medir, pero que, sin embargo, afectan las puntuaciones de prueba.

Debería quedar claro que el primer factor es deseable porque representa la verdadera cantidad del atributo en cuestión, mientras el segundo factor representa la molestia inevitable de los elementos de error que contribuyen a imprecisiones en la medición. Se puede expresar esta diferenciación conceptual en una ecuación simple:

$$X = V + e$$

donde X es la puntuación obtenida, V es la **puntuación verdadera** y e representa los **errores de medición.**

Así, estos últimos representan discrepancias entre las puntuaciones obtenidas y las puntuaciones verdaderas correspondientes:

$$e = X - V$$

Obsérvese que en las ecuaciones anteriores el error de medición e puede ser positivo o negativo. Si e es positivo, la puntuación X obtenida será más alta que la puntuación verdadera V. Por el contrario, si e es negativo, la puntuación obtenida será menor que la puntuación verdadera. Aunque es imposible eliminar todos los **errores de medición**, quienes desarrollan pruebas se esfuerzan por reducir al mínimo esta

EJEMPLO DE CASO 3-2
CONFIABILIDAD DE PRUEBA Y TESTIMONIO EN UN TRIBUNAL

La confiabilidad de los datos de prueba puede tener un impacto decisivo sobre el curso legal de las diligencias. Considérense los problemas que surgieron en un juicio de violación en donde la supuesta víctima era una mujer de 22 años con discapacidad mental. En la mayoría de los juicios de violación se acusa al inculpado de utilizar la fuerza o amenazas en la comisión de un acto sexual. Este caso era poco común porque el inculpado estaba bajo una acusación de estupro, aun cuando la víctima era mayor de edad. El abogado de la parte acusadora sostenía que, debido a su discapacidad mental, aunque la víctima había accedido, era incapaz de dar un consentimiento significativo al acto sexual.

Todos aquellos que conocían a la mujer sabían que era "lenta"; de hecho, dos profesionales de salud mental afirmaron que tenía un retraso mental leve. Con base en parte en el WAIS-R, un psicólogo consultor llegó también a un diagnóstico oficial de retraso mental. La mujer obtuvo un CI Verbal de 68 puntos, un CI de Ejecución de 75 puntos y un CI Total de 71 puntos. Un CI menor a 70 (o 75 en casos excepcionales) es sólo uno de varios criterios del retraso mental, como se analizará en el tema 7A, Pruebas con poblaciones especiales. Así, las puntuaciones se encontraban justo apenas dentro del rango de retraso mental leve y se volvieron el punto focal de las negociaciones previas entre la parte acusadora y el abogado de la defensa, quien deseaba realizar sus propios arreglos para una prueba independiente. Sin duda, el abogado de la defensa esperaba que una segunda ronda de pruebas produciría puntuaciones más altas de CI y, por tanto, refutaría la evidencia dañina de que la joven tenía retraso mental.

¿Qué podía decirle el psicólogo consultor al abogado de la acusación acerca del probable resultado de una segunda aplicación del WAIS-R? Es bien conocido que, en general, los resultados de una prueba de CI son consistentes en el corto plazo, pero ¿qué tan consistentes? ¿Era probable que la joven obtuviera puntuaciones dentro del rango de retraso mental en la segunda aplicación? La respuesta a esta última pregunta estaba llena de significado legal. Si los resultados de la segunda prueba confirmaban los datos originales, entonces la estrategia de juicio de la acusación estaba asegurada. Sin embargo, si la supuesta víctima obtenía puntuaciones sustancialmente mayores en la segunda ocasión, dichos datos arrojarían dudas acerca del diagnóstico de retraso mental. Como resultado, el fiscal se vería obligado a revisar de manera sustancial sus tácticas para el juicio —casi literalmente de un día para otro. La teoría psicométrica había dejado las alturas de la abstracción académica y había tocado tierra en el tribunal: las maniobras previas en un caso de violación dependían, en parte, de la posible repetibilidad o confiabilidad de los resultados de una prueba psicológica.

El psicólogo consultor le informó al fiscal que las probabilidades eran de 2 a 1 en favor de que el segundo CI Total cayera entre 68 y 74 puntos, suponiendo que hubiera pasado el tiempo suficiente como para eliminar cualesquiera efectos de práctica de la primera aplicación. La predicción resultó correcta. Un segundo psicólogo, contratado por la defensa, encontró que la joven tenía un CI Total de 72 puntos, sólo un punto por arriba de los datos originales. ¿Cómo fue posible pronosticar de antemano este resultado? Esta pregunta es central para el concepto de confiabilidad que se presenta en este tema. De manera incidental, al acusado se le encontró culpable de violación y se le sentenció a 15 años de prisión. Este resultado se basó, en parte, en la consistencia o confiabilidad del resultado de una prueba psicológica.

incomodidad psicométrica al prestar una atención cuidadosa a las fuentes de error de medición que se explican en la siguiente sección.

Por último, es importante destacar que la puntuación verdadera nunca se conoce. Como descubrirá el lector, se puede obtener una probabilidad de que la puntuación verdadera resida dentro de cierto intervalo y, también, se puede extraer un mejor estimado de la puntuación verdadera. Sin embargo, nunca se puede conocer con certidumbre el valor de una puntuación verdadera.

FUENTES DE ERROR DE MEDICIÓN

Como lo indica la fórmula $X = V + e$, el error de medición e es todo aquello diferente a la puntuación verdadera, que forma parte de la puntuación obtenida en la prueba. Los errores de medición pueden surgir de innumerables fuentes (Feldt y Brennan, 1989). Stanley (1971) proporciona una lista inusualmente amplia. Aquí se describirán sólo las contribuciones más importantes y probables: la selección de reactivos, la aplicación de la prueba, la calificación de la misma y los errores sistemáticos de medición.

SELECCIÓN DE REACTIVOS

Una fuente de error de medición es el instrumento en sí mismo. El autor de una prueba debe decidirse por un número finito de reactivos de una reserva potencialmente infinita de preguntas de prueba. ¿Cuáles preguntas deben incluirse? ¿Cómo deben frasearse? La selección de reactivos es crucial para la precisión de la medida.

Aunque los psicómetras se esfuerzan por obtener reactivos representativos de prueba, el conjunto particular de preguntas elegidas para una prueba podría no ser equitativo por igual para todos los individuos. Un ejemplo hipotético, deliberadamente extremo, servirá para ilustrar este punto: aun un estudiante bien preparado podría reprobar una prueba académica que enfatizara las oscuras notas de pie de página dentro del libro de texto. En contraste, un estudiante mal preparado, pero curioso, que hubiera estudiado sólo las notas de pie de página, podría tener un buen resultado en un examen de este tipo. Las calificaciones de ambas personas reflejarían cantidades masivas de error de medición. Recuérdese que en este contexto la puntuación verdadera es aquello que el estudiante realmente sabe. Para el estudiante escrupuloso, la calificación obtenida sería bastante inferior a su calificación verdadera, debido a una enorme dosis de error de medición. Para el segundo estudiante, que sabe las respuestas por casualidad, la puntuación obtenida sería bastante más alta que su puntuación verdadera, debido a un error positivo de medición.

Por supuesto, en una prueba bien diseñada, el error de medición proveniente de la muestra de reactivos será mínimo. Sin embargo, una prueba siempre constituye una muestra y nunca la totalidad del conocimiento o conducta de una persona. Como resultado, la selección de reactivos siempre es una fuente de error de medición en las pruebas psicológicas. Lo más que puede hacer un psicómetra es reducir al mínimo esta molestia indeseable al atender con cuidado los problemas relacionados con la construcción de pruebas. Se analizarán los aspectos técnicos de la selección de reactivos en el tema 4B, Construcción de pruebas.

APLICACIÓN DE LA PRUEBA

Aunque en general los examinadores proporcionan un ambiente óptimo y estandarizado de prueba, las circunstancias de aplicación pueden dar lugar a numerosas fuentes de error de medición. Ejemplos de las condiciones ambientales generales que pueden ejercer una influencia desfavorable sobre la precisión de la medición, incluyen una temperatura incómoda en la habitación, iluminación deficiente y ruido excesivo. En algunos casos, no es posible anticipar las cualidades de la situación de prueba que contribuirán al error de medición. Considérese el siguiente ejemplo: un estudiante de pregrado,

que en otros sentidos es mediocre, responde correctamente un reactivo no muy difícil de información en el que se pregunta: "¿quién escribió *Los cuentos de Canterbury*?" Cuando se le interroga después sobre si ha leído alguna obra de Chaucer, el estudiante responde: "No, pero el libro está justo detrás de usted en el librero".

Las fluctuaciones momentáneas en ansiedad, motivación, atención y nivel de fatiga en la persona examinada también pueden introducir fuentes de error de medición. Por ejemplo, una persona que no ha dormido bien la noche anterior a la prueba, podría carecer de concentración y, por tanto, leer mal las preguntas. Un estudiante distraído por una angustia emocional temporal podría responder inadvertidamente en las columnas incorrectas de la hoja de respuestas. La pesadilla clásica en este sentido es el individuo que se brinca una pregunta —digamos la número 19— pero olvida dejar en blanco el espacio correspondiente en la hoja de respuestas. Como resultado, todas las respuestas subsecuentes están desfasadas, con la respuesta 20 anotada en la hoja de respuestas como reactivo 19 y así sucesivamente.

El examinador también puede contribuir a los errores de medición en el proceso de aplicación de la prueba. En una prueba aplicada de manera oral, el hecho de que el examinador asiente de manera inconsciente con la cabeza, podría transmitirle a la persona examinada que va por buen camino, con lo cual se le guía hacia la respuesta correcta. Por el contrario, un examinador lacónico y abrupto podría intimidar a la persona, quien de otra manera estaría dispuesta a proporcionar la respuesta correcta.

CALIFICACIÓN DE LA PRUEBA

Cuando una prueba psicológica utiliza un formato diferente a aquél de opción múltiple, que se califica por medios mecánicos, se requiere cierto grado de juicio para asignar puntos a las respuestas. Por fortuna, la mayoría de las pruebas tienen criterios bien definidos para las respuestas que se dan a una de las preguntas. Estas pautas ayudan a reducir al mínimo el impacto del juicio subjetivo sobre la calificación (Gregory, 1987). Sin embargo, la subjetividad de la calificación como fuente de error de medición puede constituir un grave problema en la valoración de pruebas proyectivas o preguntas que requieren respuestas en forma de ensayo. Con respecto a las pruebas proyectivas, Nunnally (1978) señala que la persona que aplica una prueba proyectiva atraviesa por un cambio evolutivo en los criterios de calificación a medida que pasa el tiempo, llegando a considerar que un tipo particular de respuesta es cada vez más patológico con cada encuentro.

ERROR SISTEMÁTICO DE MEDICIÓN

Las fuentes de imprecisión analizadas antes se conocen de manera colectiva como **error no sistemático de medición**, lo cual implica que sus efectos sean inconsistentes e imposibles de predecir. Sin embargo, existe otro tipo de error de medición que constituye una verdadera sombra oculta en la maquinaria psicométrica. Un error sistemático de medición surge cuando, sin que el autor lo sepa, la prueba mide de manera consistente alguna otra cosa que no es el rasgo para el cual se creó. Supóngase, por ejemplo, que una escala que mide la introversión social también detecta en forma inadvertida, y de manera consistente, la ansiedad. En este caso, la ecuación que representa la relación entre puntuaciones observadas, puntuaciones verdaderas y fuentes de error de medición sería:

$$X = V + e_s + e_n$$

donde X es la puntuación obtenida, V es la puntuación verdadera, e_s es el error sistemático debido al subcomponente de ansiedad y e_n es el error no sistemático de medición descrito antes.

Dado que, por definición, su presencia no se detecta de inicio, los errores sistemáticos de medición pueden constituir un problema significativo en el desarrollo de pruebas psicológicas. Sin embargo, si los psicómetras utilizan los procedimientos apropiados de desarrollo de prueba que se analizan en el tema 4B

(Construcción de pruebas), el impacto de los errores sistemáticos de medición pueden reducirse en gran medida. A pesar de esto, los errores sistemáticos de medición sirven como recordatorio de que es muy difícil, sino es que imposible, evaluar de verdad un rasgo en aislamiento puro, sin la influencia de otros.

En esta sección se ha establecido el esquema básico de las numerosas y diversas fuentes de error de medición. El lector podría revisar el tema 2B, Proceso de Aplicación, que proporciona más detalles acerca de los múltiples factores que pueden virar el resultado de una prueba psicológica y, por tanto, introducir errores de medición.

ERROR DE MEDICIÓN Y CONFIABILIDAD

Quizá para este momento el lector se estará preguntando qué tiene que ver el error de medición con la confiabilidad. La conexión más obvia es que el error de medición reduce la confiabilidad o repetibilidad de los resultados de una prueba psicológica. De hecho, aquí se mostrará que la confiabilidad tiene una relación estadística precisa con el error de medición. La confiabilidad y el error de medición son, en realidad, sólo formas diferentes de expresar la misma preocupación: ¿qué tan consistente es una prueba psicológica? La interdependencia de ambos conceptos quedará en claro si se proporciona un bosquejo adicional de la teoría clásica de la medición.

Una suposición crucial de la teoría clásica es que los errores no sistemáticos de medición actúan como influencias aleatorias. Esto no quiere decir que las fuentes de error de medición sean totalmente misteriosas e incomprensibles en cada situación individual. En el caso de una persona, podría sospecharse que su puntuación en retención de dígitos refleja un error ligeramente negativo de medición producido por interferencia auditiva de alguien que haya tosido en el pasillo, durante la presentación del quinto reactivo. De la misma manera, podría conjeturarse que otra persona recibió el beneficio de un error positivo de medición al ver a través de un espejo, colocado detrás del examinador, la respuesta correcta al reactivo 9 en una prueba de información. Así, el error de medición no necesariamente es un acontecimiento misterioso en cada caso individual.

Sin embargo, cuando se examinan las puntuaciones de prueba de un grupo de personas, las causas del error de medición son increíblemente complejas y diversas. En este contexto, los errores no sistemáticos de medición se comportan como variables aleatorias. La teoría clásica acepta este carácter aleatorio del error de medición como una suposición axiomática.

Debido a que son sucesos aleatorios, los errores no sistemáticos de medición tienen la misma probabilidad de ser positivos o negativos y, por tanto, tendrán un promedio de cero cuando se trate de un grupo grande de sujetos. Así, una segunda suposición es que la media del error de medición es igual a cero. La teoría clásica también supone que los errores de medición no se correlacionan con las puntuaciones verdaderas. Esto hace sentido a nivel intuitivo: si las puntuaciones de error se relacionaran con otra puntuación, esto sugeriría que son sistemáticas más que aleatorias, lo cual violaría la suposición esencial de la teoría clásica. Por último, también se supone que los errores de medición no se correlacionan con errores en otras pruebas.

Se pueden resumir las características principales de la teoría clásica como sigue (Gulliksen, 1950, capítulo 2):

1. Los errores de medición son aleatorios.
2. La media del error de medición es = 0.
3. Las puntuaciones verdaderas y los errores no se correlacionan: $r_{te} = 0$.
4. Los errores en diferentes pruebas no se correlacionan $r_{12} = 0$.

Si se parte de estas suposiciones, es posible desarrollar varias implicaciones importantes para la confiabilidad y la medición. (Los siguientes puntos se basan en la suposición optimista de que los errores sistemáticos de

medición son mínimos o inexistentes para el instrumento en cuestión.) Por ejemplo, se sabe que cualquier prueba aplicada a un gran número de personas mostrará una variabilidad de las puntuaciones obtenidas que puede expresarse en términos estadísticos como una varianza, es decir, σ^2. El valor de la teoría clásica es que nos permite dividir la varianza de las puntuaciones obtenidas en dos fuentes independientes. En términos específicos, puede demostrarse que la varianza de las puntuaciones obtenidas es simplemente la varianza de las puntuaciones verdaderas más la varianza de los errores de medición:

$$\sigma_X^2 = \sigma_T^2 + \sigma_e^2$$

Referimos al lector interesado a Gulliksen (1950, capítulo 3), para los detalles sobre el cálculo.

La fórmula que se muestra antes demuestra que las puntuaciones de prueba varían como resultado de dos factores: la variabilidad en las puntuaciones verdaderas y la variabilidad debida al error de medición. La implicación obvia es que los errores de medición contribuyen a la inconsistencia de las puntuaciones obtenidas en la prueba; los resultados no continuarán siendo estables si la prueba se aplica de nuevo.

COEFICIENTE DE CONFIABILIDAD

Finalmente, se está en posición de establecer la relación precisa entre confiabilidad y error de medición. Para este momento, el lector debería haber discernido que la confiabilidad expresa la influencia relativa de las puntuaciones verdaderas y de error en las puntuaciones obtenidas en la prueba. En términos matemáticos más precisos, el **coeficiente de confiabilidad** (r_{xx}) es la proporción entre la varianza de la puntuación verdadera y la varianza total de las puntuaciones de prueba. Es decir:

$$r_{xx} = \frac{\sigma_T^2}{\sigma_X^2}$$

o de manera equivalente:

$$r_{xx} = \frac{\sigma_T^2}{\sigma_X^2 + \sigma_e^2}$$

Obsérvese que el rango de valores potenciales de r_{xx} puede obtenerse del análisis de la fórmula expresada antes. Considérese qué sucede cuando la varianza debida al error de medición (σ_e^2) es muy pequeña, cercana a cero. En ese caso, el coeficiente de confiabilidad (r_{xx}) se acerca a un valor de (σ_T^2/σ_T^2) o 1.0. En el extremo opuesto, donde la varianza debida a un error de medición es muy grande, el valor del coeficiente de confiabilidad se vuelve muy pequeño, acercándose a un límite teórico de 0.0. En suma, una prueba que tiene confiabilidad totalmente nula (con un error de medición de gran tamaño) producirá un coeficiente de confiabilidad cercano a 0, mientras que una prueba completamente confiable (sin error de medición) producirá un coeficiente de confiabilidad de 1.0. Así, el rango posible del coeficiente de confiabilidad se encuentra entre 0 y 1.0. En la práctica, todas las pruebas producen un coeficiente de confiabilidad que se coloca en algún sitio intermedio, pero entre más cercano sea el valor de r_{xx} a 1.0, mejor.

En un sentido literal, r_{xx} indica la proporción de varianza en las puntuaciones obtenidas en la prueba, que se explica por la variabilidad en las puntuaciones verdaderas. Sin embargo, la fórmula para el coeficiente de confiabilidad r_{xx} indica también una interpretación adicional. El lector recordará que las puntuaciones obtenidas se simbolizan con la letra X. De la misma manera, los subíndices en el símbolo para el coeficiente de confiabilidad significan que r_{xx} es un índice de la consistencia potencial o real de las puntuaciones obtenidas. Así, las pruebas que tienen cantidades mínimas de error de medición, producen puntuaciones consistentes y confiables; sus coeficientes de confiabilidad son cercanos a 1.0. Por el contrario, las pruebas que reflejan grandes cantidades de error de medición producen puntuaciones inconsistentes y no confiables; sus coeficientes de confiabilidad son cercanos a 0.0.

Hasta este punto, el análisis acerca de la confiabilidad ha sido conceptual más que práctico. Se ha señalado que la confiabilidad se refiere a la consistencia de la medición; que ésta se reduce en la medida en que los errores de medición dominan la puntuación obtenida; y que un índice estadístico de la confiabilidad, el coeficiente de confiabilidad, puede variar entre 0 y 1.0. Pero ¿cómo se calcula una medida estadística de la confiabilidad? Este tema se enfoca de manera indirecta, revisando primero una herramienta estadística esencial, el coeficiente de correlación. El lector descubrirá que el coeficiente de correlación, un índice numérico de la relación lineal entre dos conjuntos de puntuaciones, es una herramienta excelente para evaluar la consistencia o repetibilidad de las puntuaciones de prueba. Se proporciona una breve actualización acerca del significado de la correlación antes de proceder a un resumen de los métodos para estimar la confiabilidad.

COEFICIENTE DE CORRELACIÓN

En su aplicación más común, un **coeficiente de correlación** (r) expresa el grado de relación lineal entre dos conjuntos de puntuaciones obtenidas de las mismas personas. Los coeficientes de correlación pueden asumir valores que van de –1.00 a +1.00. Un coeficiente de correlación de +1.00 significa una relación lineal perfecta entre los dos conjuntos de puntuaciones. En particular, cuando dos medidas tienen una correlación de +1.00, la clasificación por rangos de los sujetos es idéntica en ambos conjuntos de puntuaciones. Lo que es más, cuando los datos individuales (uno de los cuales representa un par de puntuaciones de un solo sujeto) se representan en una gráfica de dispersión (figura 3–10a), se conforman a una línea perfectamente recta con una inclinación ascendente. Un coeficiente de correlación de –1.00 significa una relación igualmente fuerte, pero con una correspondencia inversa: la puntuación más alta en una variable corresponde a la puntuación más baja en la otra y viceversa. En este caso, los datos de cada individuo se conforman a una línea perfectamente recta con una inclinación descendente (figura 3–10b). Las correlaciones de +1.00 y –1.00 son muy poco comunes en la investigación psicológica y en general denotan una conclusión trivial. Por ejemplo, si en dos ocasiones en rápida sucesión contamos el número de letras en el nombre de 100 estudiantes, estos dos conjuntos de "puntuaciones" mostrarían una correlación de +1.00.

En general, las correlaciones negativas son el resultado de la manera en que se calificó a una de las dos variables. Por ejemplo, las puntuaciones en el *Category Test* (Prueba de Categorías; Reitan y Wolfson, 1993) se presentan como errores cometidos, mientras que los resultados en la Prueba de Matrices Progresivas de Raven (Raven, Court y Raven, 1983, 1986) se muestran como número de reactivos correctos. Las personas que obtienen una alta puntuación en la Prueba de Categorías (muchos errores) tendrán mayor probabilidad de obtener una baja puntuación en la prueba de Matrices Progresivas (pocos aciertos). Así, se esperaría una correlación negativa sustancial para las puntuaciones en estas dos pruebas.

Considérese la gráfica de dispersión en la figura 3–10c, que podría representar la estatura y peso hipotéticos de un grupo de personas. Como podrá ver el lector, la estatura y peso tienen una relación fuerte, pero no perfecta, entre sí. Las personas altas tienden a pesar más, las personas de baja estatura tienden a pesar menos, pero existen algunas excepciones. Si se calculara el coeficiente de correlación entre estatura y peso —una tarea estadística simple que se explica más adelante— se obtendría un valor de aproximadamente +0.80, lo cual indica una relación positiva fuerte entre estas medidas.

Cuando dos variables no tienen relación, la gráfica de dispersión asume una forma indefinida, parecida a una mancha, y el coeficiente de correlación está cercano a cero (figura 3–10d). Por ejemplo, es muy probable que en una muestra de adultos, la correlación entre tiempo de reacción y peso sea muy cercana a cero.

Por último, es importante comprender que el coeficiente de correlación es independiente

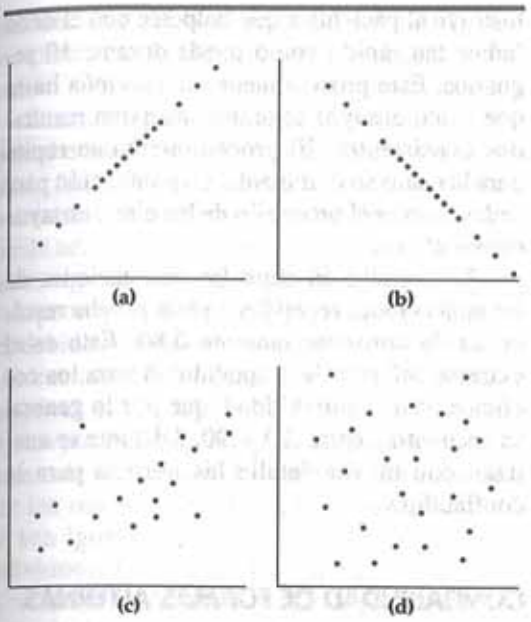


Figura 3–10. Diagrama de dispersión que representa diferentes grados de correlación.

del medio. Por ejemplo, se puede encontrar una correlación de +1.00 entre dos aplicaciones de la misma prueba, aun cuando existan diferencias medias significativas entre la preprueba y la posprueba. En resumen, la correlación perfecta no implica puntuaciones idénticas en ambas para cada individuo examinado. Sin embargo, la correlación perfecta sí significa una clasificación por rangos perfectamente ordenada de la preprueba a la posprueba, como se analizó antes.

COEFICIENTE DE CORRELACIÓN COMO COEFICIENTE DE CONFIABILIDAD

Un empleo del coeficiente de correlación consiste en estimar la consistencia de las puntuaciones en una prueba psicológica. Si los resultados de prueba son sumamente consistentes, entonces las puntuaciones de las personas que realizan la prueba en dos ocasiones se correlacionarán en gran medida e incluso se acercarán al límite teórico superior de +1.00. En este contexto, el coeficiente de correlación también es un coeficiente de confiabilidad. Aunque el cálculo de la *r* de Pearson no hace referencia a la teoría de las puntuaciones verdaderas y de error, el coeficiente de correlación sí lo hace, al reflejar la proporción de varianza en las puntuaciones obtenidas que se explica por la variabilidad en las puntuaciones verdaderas. Así, en algunos contextos, un coeficiente de correlación es un coeficiente de confiabilidad.

Este análisis introduce un método para estimar la confiabilidad de una prueba: aplicar el instrumento dos veces al mismo grupo de personas y calcular la correlación entre ambos conjuntos de puntuaciones. El enfoque *test-retest* es muy común en la evaluación de la confiabilidad, pero también existen varias otras estrategias. A medida que se revisen los métodos para estimar la confiabilidad que se describen adelante, puede ser que el lector se confunda temporalmente por la aparente diversidad de los enfoques. De hecho, los diferentes métodos caen dentro de dos amplios grupos, a saber, los enfoques de estabilidad temporal (que miden de manera directa la consistencia de las puntuaciones de prueba) y los enfoques de consistencia interna (que dependen de una sola aplicación de la prueba para estimar la confiabilidad). Debe tenerse en mente que todos los métodos eclécticos se vinculan por un tema común: la confiabilidad siempre es un intento por estimar la probable precisión o repetibilidad de las puntuaciones de prueba.

CONFIABILIDAD COMO ESTABILIDAD TEMPORAL

CONFIABILIDAD *TEST-RETEST*

El método más sencillo para determinar la confiabilidad de las puntuaciones de una prueba consiste en aplicar dos veces la misma prueba al mismo grupo de sujetos heterogéneos y representativos. Si la prueba es del todo confiable,

la segunda puntuación de cada persona se podrá pronosticar perfectamente a partir de la primera. En muchos tipos de pruebas, en particular de capacidad y de aprovechamiento, podría esperarse que, en general, los sujetos obtengan puntuaciones un tanto mejores en la segunda ocasión, debido a la práctica, maduración, escolaridad u otros efectos interventores que ocurran entre la preprueba y la posprueba. Sin embargo, en tanto la segunda puntuación se correlacione en gran medida con la primera, la existencia de los efectos de la práctica, la maduración o el tratamiento no arrojará dudas sobre la confiabilidad de una prueba psicológica.

En la figura 3–11 se presenta un ejemplo de un coeficiente de confiabilidad calculado como coeficiente de correlación **test-retest**. En este caso, se aplicó a 60 sujetos la *Finger Tapping Test* (FTT, Prueba de golpeteo con los dedos) en dos ocasiones separadas por una semana (Morrison, Gregory y Paul, 1979). La FTT, un componente de la batería de pruebas neuropsicológicas Halstead-Reitan (Reitan y Wolfson, 1985), es una medida relativamente pura de la velocidad motora. A través de un aparato estandarizado de conteo mecánico, se instruye al paciente a que golpetee con el dedo índice tan rápido como pueda durante 10 segundos. Este procedimiento se continúa hasta que cinco ensayos seguidos muestran resultados consistentes. El procedimiento se repite para la mano no dominante. La puntuación para cada mano es el promedio de los cinco ensayos consecutivos.



Figura 3–11. Diagrama de dispersión que revela un coeficiente de confiabilidad de 0.80.
Fuente: Basado en datos de Morrison, M. W., Gregory, R. J. y Paul, J. J. (1979). Reliability of the Finger Tapping Test and a note on sex differences. *Perceptual and Motor Skills*, 48, 139-142.

La correlación entre las puntuaciones de las aplicaciones repetidas de esta prueba resulta ser de aproximadamente 0.80. Éste es el extremo inferior de aceptabilidad para los coeficientes de confiabilidad, que por lo general se encuentran entre .80 y .90. Adelante se analizan con mayor detalle las normas para la confiabilidad.

CONFIABILIDAD DE FORMAS ALTERNAS

En algunos casos, los autores elaboran dos formas de la misma prueba. Estas **formas alternas** se construyen de manera independiente para satisfacer las mismas especificaciones, con frecuencia desde el nivel de cada uno de los reactivos. Así, estas formas incorporan contenido similar y cubren el mismo rango y nivel de dificultad en los reactivos, además de tener propiedades estadísticas y normativas similares. Por ejemplo, cuando se aplican de manera contrabalanceada al mismo grupo de individuos, las medias y desviaciones estándar de las formas alternas son, por lo común, bastante comparables.

Los estimados de la confiabilidad de las formas alternas se obtienen de aplicar ambas formas al mismo grupo y correlacionar los dos conjuntos de puntuaciones. Este enfoque tiene mucho en común con los métodos *test-retest* —ambas estrategias implican dos aplicaciones de prueba a las mismas personas con un intervalo intermedio. Para ambos enfoques, se esperaría que los cambios interventores en motivación y las diferencias individuales en cuanto a la cantidad de mejoría producirían fluctuaciones en las puntuaciones de prueba y, por tanto, reducirían en cierta medida los estimados de confiabilidad. Así, los estimados de

confiabilidad *test-retest* y de formas alternas comparten una considerable semejanza conceptual.

Sin embargo, existe una diferencia fundamental entre estos dos enfoques. La metodología de formas alternas introduce diferencias en la muestra de reactivos que representan una fuente adicional de varianza debida al error. Es decir, algunas personas pueden tener un mejor o un peor desempeño en una forma de la prueba, debido a la muestra particular de reactivos. Aun cuando ambas formas puedan tener el mismo nivel de dificultad, en promedio, algunos sujetos pueden encontrar que una forma es bastante más difícil (o fácil) que la otra, debido a que los reactivos que se supone son paralelos no son igualmente familiares para todos los individuos. Obsérvese que las diferencias en la muestra de reactivo no son una fuente de varianza de error en el enfoque *test-retest*, porque se utilizan reactivos idénticos en ambas aplicaciones.

Las formas alternas de una prueba también son bastante costosas —al casi duplicar los costos de publicación de una prueba y de colocación en el mercado. Debido al aumento en costo y también por las dificultades psicométricas de producir formas verdaderamente paralelas, cada vez se publican menos pruebas en este formato.

CONFIABILIDAD COMO CONSISTENCIA INTERNA

Volvemos ahora nuestra atención hacia ciertas maneras fascinantes de estimar la confiabilidad de una prueba individual sin desarrollar formas alternas y sin aplicar la prueba dos veces a las mismas personas (Feldt y Brennan, 1989). El primer enfoque correlaciona los resultados de una mitad de la prueba con los de la otra mitad y, de manera apropiada, se denomina confiabilidad de división en mitades. El segundo enfoque analiza la consistencia interna de los reactivos individuales de prueba. En este enfoque, el psicómetra busca determinar si éstos tienden a mostrar una interrelación consistente. Por último, debido a que algunas pruebas son menos que perfectamente confiables a causa de las diferencias entre calificadores, también se revisa el tema relacionado de la confiabilidad entre calificadores.

CONFIABILIDAD POR MITADES

Se obtiene un estimado de la confiabilidad **por mitades** al correlacionar los pares de puntuaciones obtenidas de mitades equivalentes de una prueba aplicada sólo una vez a una muestra representativa de examinandos. La lógica de la confiabilidad por mitades es sencilla: si las puntuaciones de las dos mitades de prueba, obtenidas de una **sola** aplicación, muestran una fuerte correlación, entonces las puntuaciones de las dos pruebas completas, obtenidas en dos aplicaciones **independientes** (el enfoque tradicional para evaluar la confiabilidad), también deberían revelar una fuerte correlación.

Es típico que los psicómetras consideren al método por mitades como complementario del enfoque de patrón oro que es el método del *test-retest*. Por ejemplo, en la estandarización del WAIS-III, la confiabilidad de la mayoría de las escalas se estableció a través del enfoque *test-retest* y del enfoque por mitades. En general, estos dos estimados de la confiabilidad son similares, aunque es frecuente que los enfoques por mitades produzcan estimados más altos de la confiabilidad.

Una justificación para el enfoque por mitades es que los problemas logísticos o de costo excesivo pueden hacer poco práctica la obtención de un segundo conjunto de puntuaciones de prueba con los mismos individuos examinados. En este caso, un estimado de la confiabilidad por mitades es lo único disponible y con seguridad es mejor que no tener ningún estimado. Otra justificación para este enfoque es que el método *test-retest* puede ser engañoso en ciertos casos.

Por ejemplo, algunas pruebas de capacidad son propensas a efectos grandes pero inconsistentes debidos a la práctica —como cuando los individuos aprenden conceptos debido a la realimentación dada como parte del procedimiento

estandarizado de prueba. Cuando los efectos de la práctica son grandes y variables, la clasificación por rango de las puntuaciones de una segunda aplicación sólo tendrá, en el mejor de los casos, una modesta asociación con la clasificación por rango de las puntuaciones de la primera aplicación. Para estos tipos de instrumentos, el coeficiente de confiabilidad *test-retest* podría ser engañosamente bajo.

Por último, los enfoques *test-retest* también pueden producir estimados engañosamente bajos de confiabilidad si el rasgo a medir es conocido por tener fluctuaciones rápidas (p. ej., ciertas medidas de estado de ánimo).

El principal desafío de la confiabilidad por mitades consiste en dividir la prueba en dos mitades aproximadamente equivalentes. En la mayor parte de las pruebas —en especial aquéllas cuyos reactivos se clasifican con base en el nivel de dificultad— la primera mitad es más fácil que la segunda. No se esperaría que los individuos obtengan puntuaciones equivalentes en estas dos porciones, de modo que este enfoque para dividir la prueba se utiliza pocas ocasiones. El método más común para obtener la división por mitades estriba en comparar las puntuaciones de los reactivos pares y nones de la prueba. Este procedimiento funciona particularmente bien cuando los reactivos se ordenan según el grado aproximado de dificultad.

Además de calcular una *r* de Pearson entre las puntuaciones de las dos mitades equivalentes de la prueba, el cálculo de un coeficiente de confiabilidad por mitades implica un paso adicional: ajustar la confiabilidad por mitades utilizando la fórmula Spearman-Brown.

FÓRMULA SPEARMAN-BROWN

Obsérvese que el método por mitades proporciona un estimado de la confiabilidad para un instrumento de la mitad de longitud que la prueba completa. Aunque existen algunas excepciones, en general una prueba más corta es menos confiable que una más larga. Esto es especialmente cierto si, en comparación con la forma más corta, la prueba más larga incorpora un contenido equivalente y de dificultad similar en los reactivos. Así, la *r* de Pearson entre las dos mitades de una prueba subestimará por lo general la confiabilidad del instrumento completo. Se necesita un método para obtener la confiabilidad de toda la prueba con base en el coeficiente de correlación por mitades.

Cuadro 3–7. Comparación de las confiabilidades por mitades y las confiabilidades Spearman-Brown correspondientes

Confiabilidad por mitades	Confiabilidad de Spearman-Brown
.5	.67
.6	.75
.7	.82
.8	.89
.9	.95

La **fórmula Spearman-Brown** proporciona el ajuste apropiado:

$$r_{SB} = \frac{2r_{mm}}{1 + r_{mm}}$$

En esta fórmula, r_{SB} es el estimado de confiabilidad de la prueba completa, calculado según el método Spearman-Brown, mientras r_{mm} es la confiabilidad por mitad de prueba. El cuadro 3–7 muestra posibles correlaciones por mitades junto con los coeficientes de confiabilidad Spearman-Brown correspondientes para toda la prueba. Por ejemplo, al utilizar la fórmula Spearman-Brown, se puede determinar que una confiabilidad por mitades de .70 es equivalente a una confiabilidad estimada para toda la prueba de .82.

CRÍTICA AL ENFOQUE POR MITADES

A pesar de que el enfoque por mitades es muy utilizado, ha recibido críticas debido a su falta de precisión:

> En lugar de proporcionar un solo coeficiente para la prueba, el procedimiento nos da diferentes coeficientes dependiendo de cuáles reactivos se agrupan cuando la prueba se divide en dos partes.

> Si una mitad puede dar un mayor coeficiente que otra, se puede tener poca fe en cualquier resultado que se obtenga de una sola mitad (Cronbach, 1951).

¿Por qué depender de una sola división? ¿Por qué no tomar un valor más típico, como la media de los coeficientes por mitades que resulten de todas las posibles divisiones de la prueba? Cronbach (1951) defendió justo ese enfoque cuando propuso una fórmula general para estimar la confiabilidad de una prueba psicológica.

COEFICIENTE ALFA

Como los propuso Cronbach (1951), y lo elaboraron de manera subsecuente otros autores (Novick y Lewis, 1967; Kaiser y Michael, 1975), el **coeficiente alfa(α)** puede considerarse como la media de todos los posibles coeficientes por mitades, corregidos según la fórmula Spearman-Brown. La fórmula para el coeficiente α es:

$$r_{\alpha} = \left(\frac{N}{N-1} \right) \left(1 - \frac{\Sigma \sigma_j^2}{\sigma^2} \right)$$

donde r_{α} es el coeficiente alfa, N es el número de reactivos, σ_j^2 es la varianza de un reactivo, $\Sigma \sigma_j^2$ es la sumatoria de las varianzas de todos los reactivos y σ^2 es la varianza de las puntuaciones totales de prueba. Como ocurre con todos los estimados de confiabilidad, el coeficiente α puede variar entre 0 y 1.00.

El coeficiente α es un índice de la consistencia interna de los reactivos; es decir, de su tendencia a correlacionarse en sentido positivo entre sí. Dado que una prueba o escala con alta consistencia interna también tenderá a mostrar estabilidad en sus puntuaciones en un enfoque *test-retest*, el coeficiente α es, por tanto, un estimado útil de la confiabilidad.

Por tradición se ha considerado que el coeficiente α es un índice de la unidimensionalidad; es decir, del grado en que una prueba o escala mide un solo factor. Los análisis recientes de Cortina (1993) y Schmitt (1996) sirven para esclarecer este error conceptual. Con seguridad el coeficiente α es un índice de la interrelación de los reactivos individuales, pero ello no es sinónimo de la unidimensionalidad de aquello que mide la prueba o escala. De hecho, es posible que una escala mida dos o más factores diferentes y, sin embargo, posea un coeficiente α muy fuerte. Schmitt (1996) da un ejemplo de una prueba de seis reactivos en la que los primeros tres se intercorrelacionan en .8, los últimos tres reactivos también tienen una correlación de .8 entre sí, mientras que las correlaciones entre los dos conjuntos de tres reactivos sólo alcanzan .3 (cuadro 3–8). Aunque ésta es, de manera irrefutable, una fuerte prueba de la existencia de dos factores, ¡el valor del coeficiente α resulta ser de .86! Para este tipo de prueba, es probable que el coeficiente α produzca una sobrestimación de la confiabilidad *test-retest*. Ésta es una de las razones por las que los psicómetras consideran a los enfoques *test-retest* como esenciales para la valoración de la confiabilidad. Con toda seguridad, el enfoque de división por mitades, en general, y el coeficiente α en particular, son abordajes valiosos a la confiabilidad, pero no pueden sustituir al sentido común del enfoque *test-retest*: cuando la misma prueba se aplica en dos ocasiones a una muestra representativa, ¿las personas evaluadas obtienen la misma asignación relativa de puntuaciones?

Cuadro 3–8. Prueba de seis reactivos con dos factores y un fuerte coeficiente α

Variable	1	2	3	4	5	6
1	—					
2	.8	—				
3	.8	.8	—			
4	.3	.3	.3	—		
5	.3	.3	.3	.8	—	
6	.3	.3	.3	.8	.8	—

Nota: coeficiente α = .86.

Fuente: Reimpreso con autorización de Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*, 350-353.

ESTIMADO KUDER-RICHARDSON DE LA CONFIABILIDAD

Cronbach (1951) ha demostrado que el coeficiente α es la aplicación general de una fórmula más específica desarrollada antes por Kuder y Richardson (1937). Su fórmula se conoce en términos generales como **fórmula Kuder-Richardson 20** o, simplemente, KR-20, en referencia al hecho de que fue la vigésima de una larga serie de derivaciones. La fórmula KR-20 es pertinente para el caso especial donde cada reactivo de prueba se califica con 0 o 1 (p. ej., incorrecto o correcto). La fórmula es:

$$KR\text{-}20 = \left(\frac{N}{N-1}\right)\left(1 - \frac{\Sigma pq}{\sigma^2}\right)$$

donde

N = número de reactivos en una prueba,
σ^2 = varianza de las puntuaciones en la prueba total,
p = proporción de examinandos que obtienen cada reactivo correcto,
q = proporción de examinandos que obtienen cada reactivo incorrecto.

El coeficiente α extiende el método Kuder-Richardson a los tipos de pruebas con reactivos que no se califican como 0 o 1. Por ejemplo, el coeficiente α puede utilizarse con una escala de actitud en la que las personas examinadas indican en cada reactivo si están muy de acuerdo, de acuerdo, en desacuerdo o muy en desacuerdo.

CONFIABILIDAD INTERCALIFICADORES

Algunas pruebas permiten una gran cantidad de juicio por parte del examinador en cuanto a la asignación de puntuaciones. Con toda seguridad, las pruebas proyectivas caen dentro de esta categoría, al igual que las pruebas de desarrollo moral y de creatividad. En la medida en que el calificador puede ser un importante factor en la confiabilidad de estos instrumentos, es imperativo informar sobre la confiabilidad intercalificadores. El cálculo de la **confiabilidad intercalificadores** es un procedimiento muy sencillo. Dos o más examinadores califican de manera independiente una muestra de las pruebas y entonces se correlacionan las puntuaciones por pares de examinadores. Por lo común, los manuales de prueba informan sobre el entrenamiento y experiencia requeridos por parte de los examinadores y después enumeran los coeficientes representativos de correlación intercalificadores.

La confiabilidad intercalificadores complementa otros estimados de confiabilidad, pero no los sustituye. De todas maneras, sería apropiado evaluar la confiabilidad *test-retest* y otros tipos de confiabilidad en una prueba que se califica de manera subjetiva. En el cuadro 3–9 se proporciona un rápido resumen acerca de los métodos para estimar la confiabilidad.

¿QUÉ TIPO DE CONFIABILIDAD ES EL APROPIADO?

Como se señaló, incluso cuando una prueba tiene sólo una forma, existen numerosos métodos para evaluar la confiabilidad: *test-retest*, por mitades, coeficiente α y métodos intercalificadores. Para las pruebas que poseen dos formas, se puede añadir un quinto método: la confiabilidad de formas alternas. ¿Cuál método es el mejor? ¿Cuándo debería utilizarse un método y no otro? Para responder a estas preguntas se necesita conocer la naturaleza y propósito de la prueba individual en cuestión.

Para las pruebas diseñadas para aplicarse a los individuos en más de una ocasión, sería razonable esperar que la prueba demostrara confiabilidad a través del tiempo —en este caso es apropiada la confiabilidad *test-retest*. Para las pruebas que afirman tener pureza factorial, el coeficiente α sería esencial. En contraste, las pruebas factorialmente complejas no darán buenos resultados con medidas de la consistencia interna. Por tanto, el coeficiente α no es un índice apropiado de la confiabilidad para todas las pruebas, sino que se aplica sólo a medidas

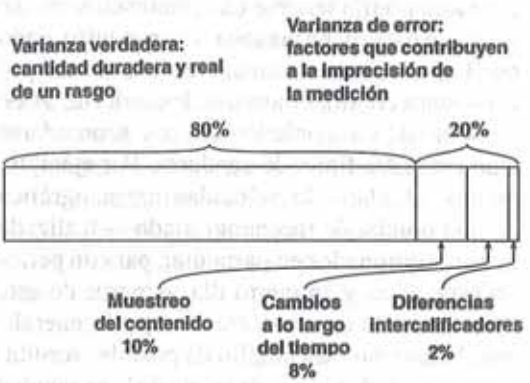
Cuadro 3–9. Breve sinopsis de los métodos para evaluar la confiabilidad

Método	Núm. de formas	Núm. de sesiones	Fuentes de varianza de error
Test-retest	1	2	Cambios a través del tiempo
Formas alternas (inmediata)	2	1	Muestra de reactivos
Formas alternas (demorada)	2	2	Muestra de reactivos Cambios a través del tiempo
División por mitades	1	1	Muestra de reactivos Naturaleza de la división
Coeficiente α	1	1	Muestra de reactivos Heterogeneidad de prueba
Intercalificadores	1	1	Diferencias entre calificadores

que están diseñadas para evaluar un solo factor. Los métodos de división en mitades funcionan bien con instrumentos cuyos reactivos se han ordenado de manera cuidadosa según su nivel de dificultad. Por supuesto, la confiabilidad intercalificadores es apropiada para cualquier prueba que requiera de subjetividad en la calificación.

Es común que los manuales de prueba señalen múltiples fuentes de información sobre la confiabilidad. Por ejemplo, el Manual del WAIS-III (Tulsky, Zhu y Ledbetter, 1997) detalla las confiabilidades de división en mitades de la mayoría de las subpruebas y también proporciona coeficientes *test-retest* para todas las subpruebas y puntuaciones de CI. El manual también cita información relacionada con la confiabilidad de formas alternas —informa de correlaciones entre el WAIS-III y su predecesor, el WAIS-R.

A fin de separar la varianza de error en sus partes componentes, será necesario calcular varios coeficientes de confiabilidad. Aunque es difícil llegar a datos precisos en el mundo real, con una base teórica se puede dividir la variabilidad de las puntuaciones en sus componentes verdaderos y de error, como se representa en la figura 3–12.



Nota: Los resultados son similares a lo que podría encontrarse si diferentes examinadores aplicaran formas alternativas de una prueba individual de inteligencia a la misma persona.

Figura 3–12. Fuentes de varianza en una prueba hipotética.

MODELO DE MUESTREO DEL DOMINIO Y LA TEORÍA DE LA CAPACIDAD PARA LA GENERALIZACIÓN

La teoría clásica resumida antes es la base para la mayoría de los análisis de confiabilidad que se citan en los informes de investigación y en

los manuales de prueba. Otro enfoque conocido como muestreo del dominio proporciona una alternativa y un concepto que induce a la reflexión acerca de los componentes de confiabilidad y error en la medición (Tryon, 1957). Aquí se analizan los aspectos esenciales de un enfoque de muestreo del dominio conocido como **teoría de la capacidad para la generalización** (Cronbach y colaboradores, 1972; Feldt y Brennan, 1989; Shavelson y Webb, 1991).

Según esta teoría, cuando los psicólogos o educadores quieren medir una variable, casi siempre tienen en mente un dominio o universo de conductas (Cronbach, 1984). Por ejemplo, cuando se evalúa la velocidad en mecanografía, el dominio o universo relacionado podría ser la velocidad promedio del individuo examinado en todo tipo de mecanografiado realizado durante un mes en particular. Si se pudiera observar y calcular la velocidad mecanográfica promedio de una persona durante todo un mes, entonces podría tenerse una puntuación en ese universo para dicha persona. Por supuesto, dado que la puntuación de un universo se refiere por lo común a enormes muestras de conducta, desde un punto de vista práctico hay que acomodarse a una muestra finita de conducta. Por ejemplo, podría calcularse la velocidad mecanográfica en una prueba de mecanografiado —realizada por un examinador en particular, para un periodo específico y en cierto día. A partir de esta muestra finita de conducta se espera generalizar al universo más amplio de posibles resultados, de aquí el nombre de teoría de la capacidad para la generalización.

Aquí son evidentes algunos paralelismos con la teoría clásica. La muestra finita de conducta en la teoría de la capacidad para la generalización es similar a la puntuación observada en la teoría clásica. Lo que es más, la puntuación del universo en aquélla es igual a la puntuación verdadera en la teoría clásica. De la misma manera en que una muestra puede o no ser una representación precisa del dominio (teoría de la capacidad para la generalización), una puntuación obtenida puede o no ser una representación precisa de la puntuación verdadera (teoría clásica). De modo que ambos enfoques tienen ciertos componentes en común.

La diferencia es que la teoría de la capacidad para la generalización reconoce que pueden existir varios universos alternativos y viables de generalización, cada uno de los cuales contiene fuentes características de error de medición. Para reconocer la existencia de universos alternativos de generalización, Cronbach y colaboradores (1972) introducen el concepto de **facetas**. Solas o en combinación, las facetas definen universos:

> Clasificamos las condiciones de observación con respecto a facetas; por ejemplo, formas de prueba, observadores y ocasiones. Esto afina la definición del universo de generalización y resalta la importancia de la definición del universo. Con frecuencia, los investigadores eligen procedimientos para evaluar la confiabilidad que definen de manera implícita un universo más estrecho de lo que requiere su teoría de sustentación. Cuando hacen esto, subestiman el "error" de medición, es decir, el error de generalización.

Por ejemplo, en un caso podría desearse conocer la manera en que una persona responde a una prueba de mecanografiado cuando se le confronta con un supervisor estricto. En esta situación, utilizar a un supervisor estricto para conducir la prueba de mecanografiado no introduce un error de medición. En otro caso, podría tenerse la intención de conocer la manera en que la persona examinada responde a una prueba de mecanografiado en un medio de oficina más típico. Para esta circunstancia, utilizar a un examinador estricto introduciría una fuente no deseada de error de medición. El elemento esencial de la teoría de la capacidad para la generalización es que no se pueden especificar las fuentes de error de medición hasta que se defina el universo relacionado de generalización.

Cronbach (1970) expresa el concepto central de esta teoría en la siguiente selección:

> Cuando utilizamos una sola observación como si representara al universo, estamos generalizando. Generalizamos acerca de los calificadores, de las selecciones a mecanografiar, quizás acerca de los días. Si las puntuaciones observadas a partir de un procedimiento concuerdan de manera estrecha con la puntuación del universo, podemos decir que la observación es "precisa",

"confiable" o "generalizable". Y dado que, entonces, las observaciones también coinciden entre sí, decimos que son "consistentes" y "tienen poca varianza de error". Poseer tantos términos causa confusión, pero no de manera tan grave. El término que se utiliza con más frecuencia en la literatura es "confiabilidad". El autor prefiere "generabilidad", porque ese término implica de inmediato "¿generalización a qué?"... Existe un grado diferente de generabilidad para cada universo. Los viejos métodos de análisis no diferencian las fuentes de variación. Tratan con una sola fuente de varianza o enredan dos o más fuentes.

La teoría clásica de la medición aún es el modelo dominante en psicometría. Sin embargo, la teoría de la capacidad para la generalización ha satisfecho una función valiosa al recordarle a los examinadores que la confiabilidad no es una cualidad estática que reside dentro de la prueba en sí; más bien, la confiabilidad es una propiedad que surge de la prueba en conjunto con las circunstancias de uso. En particular, la confiabilidad siempre depende de la manera en que se define al universo de conductas potenciales, al universo de generalización.

CIRCUNSTANCIAS ESPECIALES EN LA ESTIMACIÓN DE LA CONFIABILIDAD

Los enfoques tradicionales para la estimación de la confiabilidad pueden ser engañosos o inapropiados para algunas aplicaciones. Algunas de las situaciones más problemáticas implican características inestables, pruebas de velocidad, restricción del rango y pruebas referidas a criterio.

CARACTERÍSTICAS INESTABLES

Se supone que algunas características tienen cambios constantes en reacción ante variables situacionales o fisiológicas. Un buen ejemplo es la reactividad emocional, cuando se mide a través de la respuesta electrodérmica o galvánica de la piel. Una medida de este tipo fluctúa con rapidez en reacción ante ruidos fuertes, procesos subyacentes de pensamiento y acontecimiento ambientales estresantes. Incluso el hecho de hablar con otra persona puede despertar una respuesta electrodérmica intensa. Debido a que la verdadera cantidad de reactividad emocional cambia con tal rapidez, la prueba y la repetición de la misma deben ser casi instantáneas a fin de proporcionar un índice preciso de la confiabilidad para características inestables como la medición electrodérmica de la reactividad emocional.

PRUEBAS DE VELOCIDAD Y DE PODER

Por lo común, una **prueba de velocidad** contiene reactivos de nivel uniforme y generalmente simple de dificultad. Si el tiempo lo permite, la mayoría de los individuos deberán ser capaces de terminar con la mayor parte o con todos los reactivos de una prueba de este tipo. Sin embargo, como el nombre lo sugiere, una prueba de velocidad tiene un límite restrictivo de tiempo que garantiza que pocas personas terminen la prueba completa. Dado que los reactivos intentados tienden a ser correctos, la puntuación de una persona en una prueba de velocidad refleja en gran medida la velocidad del desempeño.

Con frecuencia, las pruebas de velocidad se contrastan con las pruebas de poder. Una **prueba de poder** concede el tiempo suficiente para que quienes realizan la prueba hagan el intento de responder a todos los reactivos, pero se construye de tal manera que no todas las personas puedan obtener una puntuación perfecta. Casi todas las pruebas contienen una mezcla de componentes de velocidad y de poder.

El elemento más importante a destacar acerca de la confiabilidad de las pruebas de velocidad es que el enfoque tradicional de división por mitades (comparación de reactivos nones y pares) producirá un coeficiente de confiabilidad engañosamente alto. Considérese a un examinando que termina 60 de los 90 reactivos en una prueba de velocidad. Es más probable que el enfoque de pares-nones muestre 30 reactivos nones correctos y 30 reactivos pares correctos. Con datos similares de otros sujetos, la correlación entre las puntuaciones en los reactivos

pares y nones se acercará necesariamente a +1.00. La confiabilidad de una prueba de velocidad debe basarse en el método *test-retest* o en la confiabilidad por mitades de dos mitades de prueba cronometradas de manera independiente. En el último caso, se requiere la corrección de Spearman-Brown.

RESTRICCIÓN DEL RANGO

La confiabilidad *test-retest* será engañosamente baja si se basa en una muestra de pacientes homogéneos, para los cuales existe una **restricción del rango** en cuanto a las características medidas. Por ejemplo, sería inapropiado estimar la confiabilidad de una prueba de inteligencia al aplicarla dos veces a una muestra de estudiantes universitarios. Este punto se ilustra con el diagrama de dispersión hipotético, pero realista, que se muestra en la figura 3–13, donde el lector puede ver una fuerte correlación *test-retest* para el rango completo de sujetos diversos, pero una correlación débil para los sujetos más inteligentes vistos de manera aislada.

CONFIABILIDAD DE LAS PRUEBAS REFERIDAS A CRITERIO

El lector recordará que en el primer tema de este capítulo se habló de que las pruebas referidas a criterio evalúan el desempeño en términos de dominio en vez de evaluar un continuo de rendimiento. Los reactivos de prueba se diseñan para identificar habilidades específicas que necesitan regularización; por tanto, los reactivos tienden a ser de la variedad "aprobado-reprobado".

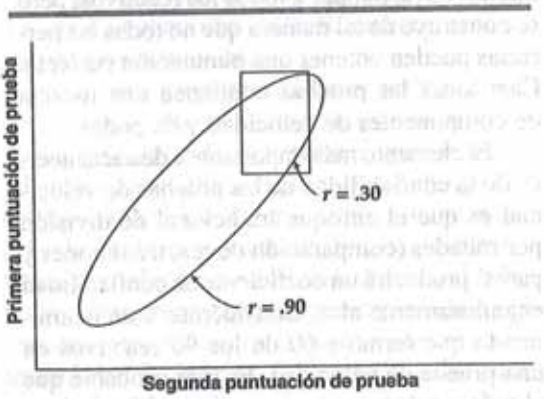


Figura 3–13. El muestreo de un rango restringido de individuos causa que la confiabilidad *test-retest* sea engañosamente baja.

La estructura de estas pruebas es de tal tipo, que la variabilidad de las puntuaciones entre los examinandos es, por lo común, bastante pequeña. De hecho, si los resultados de prueba se utilizan con propósitos de entrenamiento y los individuos continúan en entrenamiento hasta que dominan todas las habilidades requeridas en la prueba, la variabilidad en las puntuaciones se vuelve nula. Bajo tales condiciones, los enfoques tradicionales para la evaluación de la confiabilidad son simplemente inapropiados.

En el caso de muchas de las pruebas referidas a criterio, los resultados deben ser casi absolutamente precisos para que sean útiles. Por ejemplo, cualquier error de clasificación es grave si el propósito de la prueba consiste en determinar la capacidad de un individuo para manejar un automóvil con transmisión manual o "palanca de velocidades". El aspecto clave en este caso no es si las puntuaciones de *test-retest* son cercanas entre sí sino, más bien, si la clasificación ("puede hacerlo/no puede hacerlo") es la misma en ambos casos. Lo que en realidad se quiere saber es el porcentaje de personas con las cuales se llega a la misma decisión en ambas ocasiones —entre más cercano sea a 100%, mejor. Éste es sólo un ejemplo de la necesidad de técnicas especializadas en la evaluación de pruebas no normativas. Berk (1984) y Feldt y Brennan (1989) proporcionan enfoques adicionales para obtener la confiabilidad de las pruebas referidas a criterio.

INTERPRETACIÓN DE LOS COEFICIENTES DE CONFIABILIDAD

Para este momento, el lector debe ser un experto en los diferentes enfoques acerca de la confiabilidad y debe poseer al menos una idea

conceptual de cómo se calculan los coeficientes de confiabilidad. Además, se han analizado las condiciones especiales de prueba que dictan el uso de un tipo de método de confiabilidad en oposición a otro. Sin duda, el lector ha notado que aún queda por discutir una pregunta crucial: ¿cuál es un nivel aceptable de confiabilidad?

Muchos autores sugieren que la confiabilidad debe ser cuando menos de .90 para tomar decisiones sobre individuos (p. ej., Salvia e Ysseldyke, 1988; Nunnally y Bernstein, 1994). Sin embargo, en realidad no existe una respuesta definitiva para esta pregunta. Se ofrecen las pautas indefinidas que sugieren Guilford y Fruchter (1978):

> Existe cierto consenso de que para que una medida de las diferencias individuales en alguna característica sea muy precisa, la confiabilidad debe superar al .90. Sin embargo, la verdad es que muchas pruebas estándar con confiabilidades tan bajas como .70 resultan ser muy útiles. Y pruebas con confiabilidades menores que eso pueden ser útiles en investigación.

En un nivel más práctico, las normas aceptables de confiabilidad se basan en la cantidad de error de medición que el usuario puede tolerar en la aplicación sugerida para una prueba. Por fortuna, la confiabilidad y el error de medición son conceptos interdependientes. Por tanto, si el usuario de la prueba puede especificar un nivel aceptable de error de medición, entonces también es posible determinar las normas mínimas de confiabilidad que se requieren para esa aplicación específica de una prueba. Más adelante se continúa con este tema, cuando se presenta un nuevo concepto: el error estándar de medición.

CONFIABILIDAD Y ERROR ESTÁNDAR DE MEDICIÓN

Para introducir el concepto de error estándar de medición, se comenzará con un experimento imaginario. Supóngase que se pueden aplicar miles de pruebas equivalentes de CI a un individuo, y aún más, que cada sesión de prueba fuera una experiencia fresca y novedosa para un individuo cooperativo; en este experimento hipotético, la práctica y el aburrimiento no tendrían efecto sobre las posteriores calificaciones de prueba. Sin embargo, debido a los tipos de errores aleatorios que se analizan en este capítulo, las puntuaciones del desventurado paciente no serían idénticas entre sesiones de prueba, y podría tener puntuaciones un poco inferiores en una prueba debido a que estuvo despierto la noche anterior; la puntuación en otra prueba podría ser mejor debido a que los reactivos fueron peculiarmente fáciles para él. Aunque dichos factores de error son aleatorios e imposibles de predecir, de la teoría clásica de la medición se desprende que las puntuaciones obtenidas caerán dentro de una distribución normal con una media y desviación estándar precisas. Digamos que la media de las puntuaciones hipotéticas de CI para el individuo resultó ser de 110, con una desviación estándar de 2.5.

De hecho, la media de esta distribución de puntuaciones hipotéticas sería la puntuación verdadera estimada del individuo. El mejor estimado es, entonces, que este individuo tiene un CI verdadero de 110. Lo que es más, la desviación estándar de la distribución de puntuaciones obtenidas sería el **error estándar de medición** (*EEM*). Obsérvese que mientras la puntuación verdadera en la prueba tiene la probabilidad de diferir de una persona a otra, el *EEM* se considera constante, como una propiedad inherente de la prueba. Si se repite este experimento hipotético con otra persona, es probable que difiera la puntuación verdadera estimada, pero el *EEM* resultará ser de un valor similar.*

Como su nombre lo indica, el *EEM* es un índice del error de medición que atañe a la prueba en cuestión. En el caso hipotético donde el *EEM* fuera igual a cero, no habría error de medi-

* Esto sería correcto en el caso de personas de edad similar. El *EEM* puede diferir de un grupo de edad al siguiente —véase Wechsler (1997) para un ejemplo con el WAIS-III.

ción en absoluto. Entonces, la puntuación obtenida por el sujeto también sería su puntuación verdadera. Sin embargo, este resultado es simplemente imposible en una prueba realizada en la vida real. Toda prueba exhibe cierto grado de error de medición. Entre más grande sea el *EEM*, mayor será el error típico de medición. Sin embargo, la precisión o imprecisión de cualquier puntuación individual es siempre una cuestión probabilística y nunca una cantidad conocida.

Como se señaló, el *EEM* puede considerarse como la desviación estándar de las puntuaciones obtenidas por un individuo hipotético en un gran número de pruebas equivalentes, bajo la suposición de que se han descartado los efectos de la práctica y del aburrimiento. Como cualquier desviación estándar de una distribución normal, el *EEM* tiene usos estadísticos bien conocidos. Por ejemplo, 68% de las puntuaciones obtenidas caerá dentro de un *EEM* a partir de la media, al igual que 68% de los casos en una curva normal caen en una *DE* de la media.

El lector recordará que en este capítulo ya se dijo que cerca de 95% de los casos en una distribución normal caen dentro de dos *DE* de la media. Por esta razón, si el examinando realizara otra prueba más de CI, se pronosticaría con 95% de probabilidad de estar en lo correcto, que la puntuación obtenida se encontraría dentro de dos *EEM* del verdadero CI estimado de 110. Sabiendo que el *EEM* es de 2.5, se pronosticaría, por tanto, que la puntuación obtenida de CI sería de 110 ± 5, es decir, sería muy probable (95% de probabilidad) que la puntuación se encontrara entre 105 y 115.

Por desgracia, en el mundo real no se tiene acceso a las puntuaciones verdaderas y casi con toda seguridad no se pueden obtener múltiples CI de grandes números de pruebas equivalentes, como tampoco, por la misma razón, se puede tener un conocimiento directo del *EEM*. Por lo común, todo lo que se posee es un coeficiente de confiabilidad (p. ej., una correlación *test-retest* proveniente de estudios normativos) más una puntuación obtenida en la aplicación de una sola prueba. ¿Cómo puede utilizarse esta información para determinar la probable precisión de la puntuación obtenida?

CÁLCULO DEL ERROR ESTÁNDAR DE MEDICIÓN

Se ha señalado en varias ocasiones dentro del capítulo que la confiabilidad y el error de medición son conceptos interrelacionados, donde la baja confiabilidad implica un alto error de medición y viceversa. Por tanto, no debería sorprender al lector que el *EEM* se pueda calcular de manera indirecta a partir del coeficiente de confiabilidad. La fórmula es:

$$EEM = DE\sqrt{1 - r}$$

donde *DE* es la desviación estándar de las puntuaciones de prueba y *r* es el coeficiente de confiabilidad, ambos obtenidos de una muestra normativa u otro grupo grande y representativo de personas.

Se puede utilizar el CI Total del WAIS-R para ilustrar el cálculo del *EEM*. Se sabe que la *DE* del WAIS-R es aproximadamente de 15 unidades y el coeficiente de correlación es de .97 (Wechsler, 1981). Por tanto, el *EEM* para el CI Total es:

$$EEM = 15\sqrt{1 - .97}$$

que resulta ser aproximadamente de 2.5.

EEM Y PUNTUACIONES INDIVIDUALES DE PRUEBA

Considérese con cuidado qué dice el *EEM* acerca de los resultados individuales de prueba, utilizando de nuevo los CI del WAIS-R para ilustrar un concepto general. Lo que en realidad se quiere saber es la probable exactitud del CI. Si se tiene un individuo que obtiene una puntuación de 90 y la prueba se aplicó de manera competente, ¿es probable que la puntuación de CI obtenida sea exacta?

A fin de responder a esta pregunta, se necesita refrasearla. En la terminología de la teoría clásica de las pruebas, las preguntas sobre precisión en realidad implican comparaciones entre puntuaciones obtenidas y puntuaciones verdaderas. De manera específica, cuando se

cuestiona si una puntuación de CI es precisa, en realidad se está preguntando: ¿qué tan cercana está la puntuación obtenida de la puntuación verdadera?

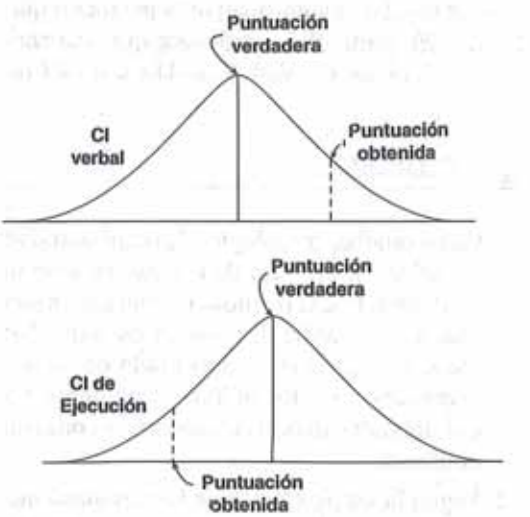
La respuesta a esta pregunta puede considerarse perturbadora a primera vista. Resulta ser que, en el caso individual, ¡nunca se puede saber con precisión qué tan cercana está la puntuación obtenida de la puntuación verdadera! A lo más que se puede llegar es a ofrecer una afirmación probabilística basada en el conocimiento de que las puntuaciones hipotéticas obtenidas para una sola persona examinada se distribuyen en forma normal con una desviación estándar igual al *EEM*. Con base en esta premisa, se sabe que la puntuación obtenida es acertada dentro de un rango de más o menos 2 *EEM*, en 95% de los casos. En otras palabras, se puede tener una seguridad de 95% de que el CI Total es preciso dentro de ± 5 puntos de CI. Este rango de más o menos cinco puntos de CI corresponde al **intervalo de confianza** de 95% para el CI Total del WAIS-R, debido a que se tiene una confianza de 95% de que la puntuación verdadera se encuentre dentro de dicho intervalo.

Los examinadores harían bien en presentar las puntuaciones de prueba en términos de un intervalo de confianza, porque esta práctica ayudaría a colocar a las puntuaciones en la perspectiva adecuada (Sattler, 1988). Un examinando que obtiene un CI de 90 debería describirse como sigue: "Mr. Doe obtuvo un CI Total de 90 puntos, que es preciso en ± 5 puntos, con una confianza de 95%". Esta redacción ayuda a advertirle a otras personas que las puntuaciones de prueba siempre incorporan cierto grado de error de medición.

EEM Y DIFERENCIAS ENTRE PUNTUACIONES

Con frecuencia se espera que los examinadores informen si una persona ha obtenido puntuaciones significativamente más altas en un área de capacidad que en otra. Por ejemplo, en general es pertinente informar si una persona tiene mayor fortaleza en tareas verbales o de ejecución o decir que no existen diferencias reales entre estas dos áreas de habilidad. El problema no es totalmente académico. Un individuo que tiene una relativa superioridad en inteligencia de ejecución podría recibir orientación para seguir una carrera práctica. En contraste, una fortaleza en inteligencia verbal podría resultar en una recomendación para perseguir intereses académicos. ¿Cómo puede determinar un examinador si una puntuación de prueba es significativamente mejor que otra?

Debe tenerse en mente que toda puntuación de prueba incorpora un error de medición. Por tanto, es posible que una persona obtuviera una puntuación verbal más alta que aquélla en ejecución cuando las puntuaciones verdaderas subyacentes —si tan sólo pudiéramos conocerlas— no revelarían diferencia ¡o incluso demostrarían el patrón contrario! (figura 3–14). La lección importante en este caso es que cuando cada una de las dos puntuaciones obtenidas refleja error de medición, la diferencia entre estas puntuaciones es bastante volátil y no debe sobreinterpretarse.



*Nota: en este caso hipotético, el CI Verbal obtenido es mayor que el CI de Ejecución obtenido, mientras que las puntuaciones verdaderas subyacentes muestran el patrón contrario.

Figura 3–14. Las puntuaciones obtenidas reflejan el error de medición y, por tanto, pueden oscurecer la relación entre las puntuaciones verdaderas.

El **error estándar de la diferencia** entre ambas puntuaciones es una medida estadística que puede ayudar al usuario de la prueba a determinar si es significativa una diferencia entre puntuaciones. El error estándar de la diferencia entre dos puntuaciones puede calcularse a partir de los *EEM* de las pruebas individuales mediante la siguiente fórmula:

$$EE_{dif} = \sqrt{(EEM_1)^2 + (EEM_2)^2}$$

donde EE_{dif} es el error estándar de la diferencia y EEM_1 y EEM_2 son los errores estándar de medición respectivos.

Se supone que las dos puntuaciones se encuentran en la misma escala o se han convertido a ella. Es decir, las pruebas deben tener la misma media y desviación estándar generales en la muestra normativa. Al sustituir $DE\sqrt{1-r_{11}}$ para el EEM_1 y $DE\sqrt{1-r_{22}}$ para EEM_2, llegamos a:

$$EE_{dif} = DE\sqrt{2 - r_{11} - r_{22}}$$

Regresamos a nuestra pregunta original para ilustrar el cálculo y uso del EE_{dif}. ¿Cómo puede determinar un examinador que una puntuación de prueba es significativamente mejor que la otra? En particular, supóngase que una persona obtiene un CI Verbal de 112 y un CI de Ejecución de 105 en el WAIS-R. ¿Siete puntos de CI representan una diferencia significativa?

Se sabe, con base en el manual del WAIS-R (Wechsler, 1981), que los CI Verbal y de Ejecución tienen cada uno desviaciones estándar de aproximadamente 15 y que sus confiabilidades respectivas son .97 y .93. El error estándar de la diferencia entre estas dos puntuaciones puede encontrarse a partir de:

$$EE_{dif} = 15\sqrt{2 - .97 - .93} = 4.74$$

Recuérdese de la discusión acerca de las distribuciones normales que 5% de los casos ocurren en los extremos, más allá de ± 1.96 desviaciones estándar. Así, las diferencias que son aproximadamente del doble del EE_{dif} (es decir, 1.96×4.74) pueden considerarse significativas en el sentido de que ocurrirán por azar sólo en 5% de las ocasiones. Por tanto, podría concluirse que es probable que las diferencias de aproximadamente nueve puntos o más entre los CI Verbal y de Ejecución reflejen diferencias reales en puntuaciones más que contribuciones aleatorias de los errores de medición. Así, es más probable que una diferencia de tan sólo siete puntos de CI no represente una diferencia significativa, auténtica, entre la inteligencia verbal y de ejecución.

RESUMEN

1. En las pruebas psicológicas, la confiabilidad se refiere al atributo de consistencia de la medición. Pocas medidas conductuales son totalmente confiables —casi en todos los casos está presente cierto grado de inconsistencia de una medición a la siguiente. La confiabilidad debería considerarse como un continuo.
2. Según la teoría clásica de las puntuaciones verdaderas y de error, cualquier puntuación de prueba refleja la influencia de dos factores: aquellos que contribuyen a la consistencia, es decir, los atributos estables que el examinador pretende medir; y aquellos que contribuyen a la inconsistencia, que incluyen las variables del sujeto, la prueba y la situación.
3. La ecuación fundamental de la teoría clásica de la medición es

 $$X = V + e$$

 donde X es la puntuación obtenida, V es la puntuación verdadera y e representa los errores de medición.
4. Los errores de medición pueden surgir durante la selección de los reactivos, la aplicación de la prueba y la calificación de la misma. Los errores sistemáticos también pueden contribuir al error de medición. Un ejemplo de error de medición producido por

la selección de reactivos es que en este proceso de selección, el autor de la prueba puede elegir reactivos que no son igualmente justos para todas las personas.

5. Los errores sistemáticos pueden surgir cuando, sin que el autor lo sepa, una prueba mide alguna otra cosa además del rasgo que se pretende medir. Por ejemplo, una prueba para medir la introversión social puede evaluar la ansiedad de manera consistente e inadvertida.
6. Las suposiciones básicas de la teoría clásica de la medición son: 1) los errores de medición son aleatorios, 2) la media del error de medición es igual a cero, 3) las puntuaciones verdaderas y las puntuaciones de error no tienen correlación y, 4) los errores en diferentes pruebas no tienen correlación. De estas suposiciones se deduce que la varianza de las puntuaciones obtenidas es simplemente la varianza de las puntuaciones verdaderas más la varianza de los errores de medición.
7. La confiabilidad expresa la influencia relativa de las puntuaciones verdaderas y de error sobre las puntuaciones obtenidas en la prueba. El coeficiente de confiabilidad es la proporción de la varianza de las puntuaciones verdaderas con respecto a la varianza total de las puntuaciones de prueba (varianza de las puntuaciones verdaderas más varianza de las puntuaciones de error). El valor del coeficiente de confiabilidad puede variar entre 0 y 1.0.
8. El coeficiente de correlación producto-momento de Pearson puede utilizarse para estimar la consistencia de las puntuaciones de una prueba psicológica. Esta forma de confiabilidad se conoce como confiabilidad *test-retest*. La confiabilidad de formas alternas se calcula a través de correlacionar las puntuaciones en dos formas equivalentes, aplicadas en forma contrabalanceada a un amplio grupo de sujetos heterogéneos.
9. Los enfoques de confiabilidad relacionados con la consistencia interna incluyen la confiabilidad por mitades, en la cual se correlacionan entre sí las puntuaciones de ambas mitades de una prueba, y el coeficiente alfa (α), que puede considerarse como la media de todos los posibles coeficientes por mitades.
10. Para las pruebas que requieren del juicio del examinador para la asignación de puntuaciones, es necesaria la confiabilidad intercalificadores. El cálculo de la confiabilidad intercalificadores es sencillo: dos o más examinadores califican de manera independiente una muestra de pruebas y después se correlacionan las puntuaciones por pares de examinadores.
11. La teoría de la capacidad para la generalización, un modelo de medición relacionado con el muestreo del dominio, constituye un punto de vista alternativo para la comprensión de la confiabilidad. Según este modelo puede especificarse un universo relacionado de conductas, llamado dominio, para cualquier tarea de medición. Una prueba constituye una muestra de conducta que puede o no ser representativa del dominio.
12. Los enfoques tradicionales para la estimación de la confiabilidad pueden ser engañosos o inapropiados para las siguientes aplicaciones: cuando la característica medida es sumamente volátil o inestable; en el caso de pruebas de velocidad que tienen reactivos de dificultad menor; y cuando los sujetos son sumamente homogéneos en cuanto a la característica medida.
13. En el caso de las pruebas referidas a criterio, los resultados deben ser casi perfectamente confiables como para considerarlos útiles. Dado que, con frecuencia estas pruebas tienen una cualidad de "puede hacerlo/no puede hacerlo", la repetibilidad de la clasificación es un método para evaluar su confiabilidad.
14. La confiabilidad se relaciona de manera inversa con el error estándar de medición (*EEM*), el cual determina el intervalo de confianza que rodea a la puntuación de todo examinando. Por ejemplo, el intervalo de confianza de 95% se encuentra a $\pm$ 2 *EEM* de la puntuación obtenida por el individuo.

TÉRMINOS Y CONCEPTOS CLAVE

Confiabilidad p. 94
Teoría clásica de medición p. 94
Puntuación verdadera p. 94
Errores de medición p. 94
Error no sistemático de medición p. 97
Coeficiente de confiabilidad p. 99
Coeficiente de correlación p. 100
Confiabilidad *test-retest* p. 101–102
Confiabilidad de formas alternas p. 102
Confiabilidad por mitades p. 103
Fórmula Spearman-Brown p. 104
Coeficiente alfa (α) p. 105
Fórmula Kuder-Richardson 20 p. 106
Confiabilidad intercalificadores p. 106
Teoría de la capacidad para la generalización p. 108
Prueba de velocidad p. 109
Prueba de poder p. 109–110
Restricción del rango p. 110
Error estándar de medición p. 111
Intervalo de confianza p. 113
Error estándar de la diferencia p. 114 ■

CAPÍTULO 4

Validez y desarrollo de pruebas

TEMA 4A
Conceptos básicos de validez

Validez: definición
Ejemplo de caso 4–1 La investigación como demostración esencial de la validez de una prueba
Validez de contenido
Validez relacionada con el criterio
Validez de constructo
Enfoques a la validez de constructo
Preocupaciones externas a la validez y ampliación del espectro de la validez de una prueba
Resumen
Términos y conceptos clave

Como sabe casi cualquier estudiante de psicología, el mérito de una prueba psicológica se determina, en primer lugar, por su confiabilidad, pero en última instancia por su validez. En el capítulo anterior se señaló que la confiabilidad puede estimarse a través de muchos métodos, en apariencia diferentes, que van desde el enfoque conceptualmente directo del *test-retest* hasta las metodologías teóricamente más complejas de la consistencia interna. Sin embargo, sin importar el método que se utilice, la evaluación de la confiabilidad se reduce de modo invariable a una simple estadística de resumen, el coeficiente de confiabilidad. En este capítulo se investiga el aspecto más difícil y complejo de la validez —el significado de una puntuación de prueba. El concepto de validez todavía se encuentra en evolución y, por tanto, provoca una cantidad mucho mayor de controversias que su humilde y formal pariente, la confiabilidad (AERA, APA y NCE, 1999). En el tema 4A, Conceptos básicos de validez, se presentan los

conceptos esenciales acerca de la validez, incluyendo la división tripartita estándar en validez de contenido, relacionada con el criterio y de constructo. También se analizan las cuestiones externas a la validez, que incluyen los efectos colaterales y consecuencias involuntarias de la aplicación de pruebas. Estas cuestiones han fomentado una definición más amplia de la validez de prueba que se extiende más allá de los conceptos técnicos de contenido, criterios y constructos. En el tema 4B, Construcción de pruebas, se destaca que la validez debe formar parte integral de la prueba desde el principio, más que limitarse a las etapas finales del desarrollo de prueba.

Dicho en términos sencillos, la validez de una prueba es el grado en que mide aquello que afirma medir. Durante mucho tiempo, los psicómetras han reconocido que la validez es la característica más fundamental e importante de una prueba. Después de todo, este término define el **significado** de las puntuaciones de prueba. La confiabilidad es importante también, pero sólo hasta el punto donde restrinja la validez. En la medida en que una prueba sea poco confiable, no puede ser válida. Esto puede expresarse desde una perspectiva alterna: la confiabilidad es precursor necesario, pero no suficiente, de la validez.

Los autores de pruebas tienen la responsabilidad de demostrar que los nuevos instrumentos satisfacen los propósitos para los cuales se diseñaron. Sin embargo, a diferencia de la confiabilidad de prueba, la validez no es una cuestión simple que se resuelve con facilidad a través de unos cuantos estudios rudimentarios. La validación es un proceso en evolución, que comienza con la construcción de prueba y continúa de manera indefinida:

> Después de que se pone en operación una prueba, puede continuarse afinando, refinando y enriqueciendo el significado interpretativo de sus puntuaciones a través de la acumulación gradual de observaciones clínicas y de proyectos especiales de investigación... La validez de prueba es algo viviente; no muere y se le embalsama en el momento en que se publica la prueba (Anastasi, 1986).

La validez de una prueba depende de la acumulación de descubrimientos de investigación (ejemplo de caso 4–1). En las siguientes secciones se examinan los tipos de evidencia que se buscan en la validación de una prueba psicológica.

VALIDEZ: DEFINICIÓN

Se comienza con una definición de **validez**, parafraseada del influyente *Standards for Educational and Psychological Testing* (AERA, APA y NCME, 1985, 1999):

> Una prueba es válida al grado en que las inferencias que se realicen a partir de ella sean apropiadas, significativas y útiles.

Véase que una puntuación de prueba por sí misma carece de significado hasta que el examinador extrae inferencias de ella con base en el manual de prueba o en otras conclusiones de la investigación. Por ejemplo, el conocimiento de que una persona evaluada ha obtenido una puntuación un tanto elevada en la escala de Depresión del MMPI-2 no es de utilidad particular. Este resultado sólo se vuelve valioso en el momento en que el examinador infiere características conductuales a partir de él. Con base en la investigación existente, el examinador podría concluir que "la puntuación elevada en Depresión sugiere que la persona examinada tiene poca energía y una actitud pesimista ante la vida". La escala de Depresión del MMPI-2 posee validez psicométrica en la medida en que tales inferencias sean apropiadas, significativas y útiles.

Por desgracia, en pocas ocasiones es posible resumir la validez de una prueba en términos de una estadística única y precisa. La determinación de la idoneidad, significado y utilidad de las inferencias, requiere por lo común de numerosos estudios acerca de las relaciones entre el desempeño en una prueba y otras conductas observadas de manera independiente. La validez refleja un juicio en evolución,

EJEMPLO DE CASO 4–1
LA INVESTIGACIÓN COMO DEMOSTRACIÓN ESENCIAL DE LA VALIDEZ DE UNA PRUEBA

En colaboración con su asesor de la facultad, un estudiante de posgrado en psicología intentó diseñar una medida de lápiz y papel acerca de la timidez. Éste siguió los pasos apropiados de selección de reactivos y construcción de escala que se analizan en la siguiente sección (tema 4B, Construcción de pruebas) y después trató de validar su instrumento correlacionándolo con medidas existentes, como las 10 escalas clínicas del MMPI. Por desgracia, esta nueva medida, que de manera tentativa se llamó Índice de timidez, resultó tener una correlación muy elevada con la escala de Depresión del MMPI ($r = .65$) y, en esencia, ninguna correlación con la Escala de Introversión Social del MMPI ($r = .08$). La medida de **timidez** del estudiante de posgrado era, en apariencia, un índice de ineptitud social y de la depresión resultante. Existe una moraleja para esta breve historia: el rasgo medido por una prueba específica se define a través de la investigación acerca de su validez y no por el nombre que el autor decida ponerle al instrumento.

basado en la investigación, acerca de qué tan adecuadamente mide la prueba aquel atributo para el cual se diseñó. En consecuencia, la validez de las pruebas no puede captarse con facilidad a través de resúmenes estadísticos exactos, sino que más bien se le puede caracterizar en un continuo que va de **débil** a **aceptable** a **fuerte.**

Tradicionalmente, las diferentes maneras de acumular evidencia de la validez se han agrupado en tres categorías:

- Validez de contenido.
- Validez relacionada con el criterio.
- Validez de constructo.

En breve se profundizará sobre esta perspectiva tripartita de la validez, pero antes deben exponerse ciertas reservas. El empleo de estas etiquetas convenientes no implica que existan tipos precisos de validez o que un procedimiento específico de validación sea mejor para utilizarse en una prueba y no en otra:

> Una validación ideal incluye varios tipos de evidencia, que abarcan las tres categorías tradicionales. Si no intervienen otros factores, más fuentes de evidencia son mejores que pocas. Sin embargo, la calidad de la evidencia es de importancia primordial y una sola línea sólida es preferible a numerosas de calidad cuestionable. El juicio profesional debe guiar las decisiones concernientes a las formas de evidencia que son más necesarias y factibles en vista de los usos propuestos para la prueba y de cualesquiera alternativas probables para la misma (AERA, APA y NCME, 1985).

Estos puntos pueden resumirse al destacar que la validez es un concepto unitario que se determina por el grado en que una prueba mide lo que pretende medir. Las inferencias que se extraen de una prueba válida son apropiadas, significativas y útiles. En este sentido, debería ser aparente que casi cualquier estudio empírico que relacione las puntuaciones de prueba con otros datos es una fuente potencial de información de validez (Anastasi, 1986; Messick, 1995).

VALIDEZ DE CONTENIDO

La **validez de contenido** se determina por el grado en que las preguntas, tareas o reactivos de la prueba constituyen una muestra representativa del universo de conductas para el cual se diseñaron; en teoría, no es más que una cuestión de muestreo (Bausell, 1986). Los reactivos de una prueba pueden considerarse como una muestra obtenida de una población más amplia

de reactivos potenciales que definen aquello que el investigador realmente desea medir. Si la muestra (reactivos específicos de la prueba) es representativa de la población (todos los reactivos posibles), entonces la prueba posee validez de contenido.

Dicha validez es un concepto útil cuando se sabe mucho acerca de la variable que el investigador desea medir. En particular en el caso de las pruebas de rendimiento, con frecuencia es posible especificar de antemano el universo relacionado de conductas. Por ejemplo, cuando se crea una prueba de rendimiento para ortografía, un investigador puede identificar casi todas las palabras posibles que los niños de tercer grado deberían saber. La validez de contenido de esta prueba se garantizaría, en parte, si se tomara una muestra aleatoria de palabras de diverso nivel de dificultad, de una lista preexistente.

Sin embargo, los autores de pruebas deben tener cuidado de especificar también el universo relacionado de respuestas. Con demasiada frecuencia se da por sentado el formato de opción múltiple:

> Si el constructor conserva una mente abierta en cuanto a sus fines, con frecuencia decidirá que la tarea debería requerir de una respuesta elaborada por el estudiante —respuestas abiertas o, si se han de reducir al mínimo las inhibiciones, respuestas orales. Tampoco se han de descuidar las instrucciones que se den al individuo y el entorno social de la prueba cuando se define la tarea (Cronbach, 1971).

Con respecto al rendimiento en ortografía, no puede suponerse que una prueba de opción múltiple medirá las mismas habilidades en ortografía que una prueba oral o que el cálculo de la frecuencia de errores ortográficos en una composición escrita. Así, cuando se evalúa la validez de contenido, la especificación de la respuesta también es una parte integral de la definición del universo relacionado de conductas.

La validez de contenido es más difícil de garantizar cuando la prueba mide un rasgo definido de modo aproximado. ¿Qué esperanza puede tener un autor de pruebas de especificar el universo de reactivos potenciales para una medida de ansiedad? En estos casos, cuando el rasgo a medir es menos tangible, ningún autor de pruebas en pleno uso de sus facultades intentaría construir el universo literal de reactivos potenciales de prueba. En vez de ello, lo que por lo general se toma como validez de contenido es la opinión considerada de jueces expertos. En efecto, el autor de pruebas afirma que "un panel de expertos revisó de manera cuidadosa la especificación del dominio y juzgó que las siguientes preguntas de prueba poseen validez de contenido". La figura 4–1 reproduce una muestra de un formato de calificación para jueces que se emplea para determinar la validez de contenido de las preguntas de prueba.

CUANTIFICACIÓN DE LA VALIDEZ DE CONTENIDO

Lawshe (1975), Martuza (1977) y otros autores han analizado los métodos estadísticos para determinar la validez general de contenido de una prueba a partir de los juicios de expertos. Estos métodos tienden a ser muy especializados y no han obtenido una amplia aceptación; sin embargo, sus enfoques pueden servir como un modelo para el punto de vista de sentido común acerca de la concordancia intercalificadores como base para la validez de contenido.

Cuando dos jueces expertos valoran los reactivos individuales de una prueba en la escala de cuatro puntos que se propone en la figura 4–1, las calificaciones de cada juez para cada reactivo pueden dividirse en dos tipos: relevancia débil (calificaciones de 1 o 2) contra relevancia fuerte (calificaciones de 3 o 4). Entonces, pueden anotarse las calificaciones conjuntas de ambos jueces para cada reactivo en el cuadro de concordancia de dos por dos que se presenta en la figura 4–2. Por ejemplo, si los dos jueces consideraron que un reactivo era sumamente relevante (relevancia fuerte), se le colocaría en la casilla D. Si el primer juez consideró que un reactivo era muy relevante (relevancia fuerte) pero el segundo lo consideró sólo como apenas relevante (relevancia débil), el reactivo se colocaría en la casilla B.

Revisor: _____ Fecha: _____

Por favor lea cuidadosamente la especificación del dominio para esta prueba. A continuación, indique por favor qué tan bien siente usted que cada reactivo refleja la especificación del dominio. Juzgue a cada uno de manera individual con base en la comparación entre su contenido y aquel definido por la especificación del dominio. Por favor utilice la escala de cuatro puntos que se muestra abajo:

1	2	3	4
no relacionado	un tanto relacionado	bastante relacionado	muy relacionado

Figura 4–1. Muestra de formato de calificación de reactivos por parte de jueces para la determinación de la validez de contenido.
Fuente: Basado en Martuza (1977), Hambleton (1984), Bausell (1986).

Obsérvese que la casilla D es la única que refleja concordancia válida entre los jueces; las otras, implican desacuerdo (casillas B y C) o concordancia en que un reactivo no debe estar en la prueba (casilla A). En la figura 4–3 se presentan los resultados hipotéticos para una prueba de 100 reactivos. Con la siguiente fórmula se puede obtener un coeficiente de la validez de contenido:

$$\text{validez de contenido} = \frac{D}{(A + B + C + D)}$$

Por ejemplo, en la prueba de 100 reactivos, ambos jueces concordaron en que 87 reactivos tenían una fuerte relevancia (casilla D), de modo que el coeficiente de validez de contenido sería 87/(4+4+5+87) o .87. Si se utilizan más de dos jueces, este procedimiento de cálculo podría llevarse a cabo con todas las posibles combinaciones de pares de jueces e informar el coeficiente promedio. Una nota importante: un coeficiente de validez de contenido es sólo un trozo de evidencia en la valoración de una prueba. Dicho coeficiente no establece en sí mismo la validez de la prueba.

El enfoque de sentido común para la validez de contenido que se propone aquí sirve bien como mecanismo de señalamiento que ayuda a seleccionar y eliminar los reactivos existentes que los calificadores expertos consideren inapropiados. Sin embargo, no puede identificar aquéllos inexistentes que deberían añadirse

		JUEZ EXPERTO # 1: Relación débil (reactivo calificado con 1 o 2)	JUEZ EXPERTO # 1: Relación fuerte (reactivo calificado con 3 o 4)
JUEZ EXPERTO # 2	Relación débil (reactivo calificado con 1 o 2)	A	B
	Relación fuerte (reactivo calificado con 3 o 4)	C	D

Figura 4–2. Modelo de concordancia intercalificadores para la validez de contenido.

		JUEZ EXPERTO # 1: Relación débil (reactivo calificado con 1 o 2)	JUEZ EXPERTO # 1: Relación fuerte (reactivo calificado con 3 o 4)
JUEZ EXPERTO # 2	Relación débil (reactivo calificado con 1 o 2)	4 reactivos	5 reactivos
	Relación fuerte (reactivo calificado con 3 o 4)	4 reactivos	87 reactivos

Figura 4–3. Ejemplo hipotético del Modelo de concordancia para la validez de contenido de una prueba de 100 reactivos.

a una prueba para ayudar a constituir una reserva de preguntas más representativas del dominio propuesto. Una prueba puede tener un fuerte coeficiente de validez de contenido y aun así carecer de sentidos más sutiles. La cuantificación de la validez de contenido no sustituye a una selección cuidadosa de reactivos.

VALIDEZ DE FACIE

Aquí se hará un breve paréntesis para mencionar la validez de Facie, que en realidad no es, de ninguna manera, un tipo de validez. No obstante, el concepto aparece cuando se trata de pruebas y, por tanto, requiere de una breve explicación. Una prueba tiene **validez de Facie** si parece válida para los usuarios, examinadores y, en especial, para las personas evaluadas. Esta validez es en realidad una cuestión de aceptabilidad social y no una forma técnica de validez en la misma categoría que las de contenido, relacionada con el criterio o de constructo (Nevo, 1985). Desde un punto de vista de relaciones públicas, es crucial que la prueba posea validez de Facie —de otra manera aquellos que realicen la prueba se sentirían insatisfechos y dudosos acerca del valor de la prueba psicológica. Sin embargo, la validez de Facie no debe confundirse con la objetiva, que se determina a través de la relación de las puntuaciones de prueba con otras fuentes de información. De hecho, una prueba puede poseer una validez de Facie extremadamente fuerte —los reactivos podrían parecer relevantes en alto grado para aquello que supuestamente mide el instrumento— y, sin embargo, producir puntuaciones que carecen de significado y que no tienen ningún tipo de utilidad predictiva.

VALIDEZ RELACIONADA CON EL CRITERIO

La **validez relacionada con el criterio** se demuestra cuando se confirma que una prueba es eficaz para estimar el desempeño del examinando en alguna medida de resultado. En este contexto, la variable de interés principal es la medida de resultado, llamada **criterio**. La puntuación de prueba es útil sólo en la medida en que proporcione una base para la predicción precisa del criterio. Por ejemplo, un examen de ingreso a la universidad que sea razonablemente preciso en la predicción del promedio de calificaciones escolares subsecuentes de la persona evaluada, poseería validez relacionada con el criterio.

Dos enfoques diferentes a la evidencia de validez se engloban bajo el encabezado de validez relacionada con el criterio. En la **validez concurrente**, las medidas de criterio se obtienen aproximadamente al mismo tiempo que las puntuaciones de prueba. Por ejemplo, el diagnóstico psiquiátrico actual de los pacientes sería una medida apropiada de criterio para proporcionar evidencia de validez para una prueba psicodiagnóstica de lápiz y papel. En la **validez predictiva**, las medidas de criterio se obtienen en el futuro, en general meses o años después de que se obtienen las puntuaciones de prueba, como en el caso de calificaciones universitarias pronosticadas a partir de un examen de ingreso. Cada uno de estos dos enfoques es más adecuado para las diferentes situaciones de prueba que se analizan en secciones subsecuentes. No obstante, antes de revisar la naturaleza de la validez concurrente y predictiva, se debe examinar una pregunta fundamental: ¿cuáles son las características de un buen criterio?

CARACTERÍSTICAS DE UN BUEN CRITERIO

Como se señaló, un **criterio** es cualquier medida de resultado contra la cual se valida una prueba. En términos prácticos, un criterio puede ser casi cualquier cosa. Algunos ejemplos ayudarán a ilustrar la diversidad de los criterios posibles. Una prueba de habilidad de manejo con base en un simulador podría validarse contra un criterio de "número de multas de tránsito recibidas en los últimos 12 meses". Una escala

que mide readaptación social podría validarse contra un criterio de "número de días de internamiento en un hospital psiquiátrico en los últimos tres años". Una prueba de potencial de ventas podría validarse contra un criterio de "valor en dólares de los bienes vendidos durante el año anterior". El ingenio del autor de la prueba determina, en parte, la elección de los criterios. Sin embargo, los criterios deben ser más que sólo imaginativos, también deben ser confiables, apropiados y estar libres de contaminación de la prueba en sí.

El criterio debe ser confiable en sí mismo, si es que ha de ser un índice útil de aquello que mide la prueba. Como recordará el lector, el significado de confiabilidad es la consistencia de las puntuaciones, por lo que la necesidad de una medida confiable de criterio es intuitivamente obvia. Después de todo, poco confiable significa poco predecible. Un criterio poco confiable será intrínsecamente poco predecible, muy aparte de los méritos de la prueba.

Considérese el caso en que las puntuaciones de un examen de ingreso a la universidad (la prueba) se utilizan para predecir el promedio de calificaciones subsecuente (el criterio). La validez del examen de ingreso puede estudiarse al calcular la correlación (r_{xy}) entre las puntuaciones del examen de ingreso y el promedio de calificaciones de una muestra representativa de estudiantes. Para propósitos de un estudio de validez, sería ideal si a los estudiantes se les concediera la inscripción abierta o sin selección, de modo que se evitara una restricción del rango en la variable de criterio. En cualquier caso, el coeficiente de correlación resultante se denomina **coeficiente de validez.***

El límite superior teórico del coeficiente de validez se restringe en función de la confiabilidad tanto de la prueba como del criterio:

$$r_{xy} = \sqrt{(r_{xx})(r_{yy})}$$

El coeficiente de validez siempre es menor o igual a la raíz cuadrada de la confiabilidad de la prueba multiplicada por la confiabilidad del criterio. En otras palabras, en la medida en que la confiabilidad de la prueba o del criterio (o de ambos) sea baja, también se reducirá el coeficiente de validez. Si se regresa al ejemplo de un examen de ingreso que se utiliza para pronosticar el promedio de calificaciones en la universidad, debe concluirse que el coeficiente de validez para una prueba de este tipo siempre será menor a +1.00, lo cual se debe en parte a la falta de confiabilidad de las calificaciones en la universidad, y también en parte a la falta de confiabilidad de la prueba en sí.

Una medida de criterio también debe ser apropiada para la prueba bajo investigación. El libro de consulta de *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1985) incorpora este importante asunto como una norma separada:

> Todas las medidas de criterio deben describirse de manera precisa y hacerse explícita la fundamentación para elegirlas como criterios relacionados.

Por ejemplo, en el caso de las pruebas de interés, en ocasiones es poco claro si la medida de criterio debe indicar satisfacción, éxito o continuidad de las actividades en cuestión. La elección entre estas variantes sutiles en el criterio debe hacerse con cuidado, con base en un análisis de qué es lo que la prueba de interés pretende medir.

Un criterio también debe estar libre de contaminación de la prueba en sí. Lehman (1978) ha ilustrado este punto en un estudio sobre la validez relacionada con el criterio de una medida del cambio vital. El *Schedule of Recent Events* (Inventario de Acontecimientos Recientes. SRE, Holmes y Rahe, 1967) es un instrumento utilizado de manera amplia que proporciona un índice cuantitativo de las acumulaciones de acontecimientos vitales estresantes (p. ej., divorcio, ascenso en el trabajo, multas de tráfico). Las calificaciones en el SRE se correlacionan de manera modesta con medidas de criterio tales como enfermedad física y alteración psicológica.

* De manera propositiva no nos referimos a dicha estadística como el coeficiente de validez. Recuérdese que ésta es un concepto unitario que se determina a través de múltiples fuentes de información que pueden incluir la correlación entre prueba y criterio.

No obstante, muchas medidas de criterio aparentemente apropiadas incorporan reactivos similares o idénticos a los del SRE. Por ejemplo, con frecuencia las pruebas de detección de síntomas psiquiátricos investigan cambios en alimentación, sueño o actividades sociales. Por desgracia, el SRE incorpora preguntas que investigan lo siguiente:

cambio en hábitos de alimentación
cambio en hábitos de sueño
cambio en actividades sociales

Si la prueba de detección contiene los mismos reactivos que el SRE, entonces la correlación entre estas dos medidas se inflará de manera artificial. Esta probable fuente de error en la validación de la prueba se conoce como **contaminación del criterio**, dado que éste se "contamina" por su parecido artificial con la prueba.

La contaminación del criterio también es posible cuando éste consiste en calificaciones de expertos. Si éstos también tienen conocimiento de las puntuaciones de prueba de las personas examinadas, esta información puede influir (de manera consciente o inconsciente) en sus calificaciones. Cuando se valida una prueba contra un criterio de calificaciones de expertos, las puntuaciones de prueba deben mantenerse en la más estricta reserva hasta que se hayan obtenido las calificaciones.

Ahora que el lector conoce las características generales de un buen criterio, se revisará la aplicación de este conocimiento al análisis de la validez concurrente y predictiva.

VALIDEZ CONCURRENTE

En un estudio de validez concurrente, las puntuaciones de prueba y la información de criterio se obtienen de manera simultánea. En general, la evidencia concurrente de la validez de una prueba es deseable en el caso de pruebas de rendimiento, aquellas que se utilizan para otorgar licencias o certificación y de diagnóstico clínico. Una evaluación de la validez concurrente indica el grado en que las puntuaciones de una prueba estiman de manera precisa la posición actual de un individuo en el criterio relacionado. Por ejemplo, una prueba de rendimiento en aritmética tendría validez concurrente si sus puntuaciones pudieran utilizarse para predecir, con una precisión razonable, la posición actual de los estudiantes en un curso de matemáticas. Un inventario de personalidad tendría validez concurrente si las clasificaciones diagnósticas que se obtienen de él concuerdan de manera aproximada con las opiniones de psiquiatras o psicólogos clínicos.

Una prueba con validez concurrente demostrada proporciona un atajo para obtener información que de otra manera podría requerir de una inversión considerable de tiempo profesional. Por ejemplo, el procedimiento de asignación de casos en una clínica de salud mental puede agilizarse si se utiliza una prueba con validez concurrente demostrada para tomar las decisiones iniciales de detección. De este modo, los pacientes con trastornos graves que requieren de un estudio clínico y tratamiento intensivo inmediato, se identificarían con rapidez mediante una prueba de lápiz y papel. Por supuesto, las pruebas no tienen el propósito de reemplazar a los especialistas en salud mental, pero pueden ahorrar tiempo en las fases iniciales de diagnóstico.

Es frecuente que las correlaciones entre una nueva prueba y otras existentes se citen como evidencia de validez concurrente. Ésta es un arma de dos filos —que las pruebas viejas validen una nueva— sin embargo, es apropiado si se satisfacen dos condiciones. Primero, las pruebas criterio (existentes) deben haberse validado a través de correlaciones con datos conductuales apropiados que no se hayan obtenido con pruebas. En otras palabras, la red de relaciones entrecruzadas debe hacer tierra en algún punto con una conducta de la vida real. En segundo lugar, el instrumento a validar debe medir el mismo constructo que las pruebas criterio. Así, es totalmente apropiado que los autores de una nueva prueba de inteligencia informen de correlaciones entre ella y pilares tales como las escalas Stanford-Binet y Wechsler.

VALIDEZ PREDICTIVA

En un estudio de validez predictiva, las puntuaciones de una prueba se utilizan para estimar medidas de resultado obtenidas en fecha posterior. La validez predictiva es de particular relevancia para exámenes de ingreso y pruebas de empleo. Dichas pruebas comparten una función común —determinar quién está en probabilidad de tener éxito en una tarea futura. Un criterio relacionado para un examen de ingreso a la universidad sería el promedio de calificaciones del primer año, mientras que una prueba de empleo podría validarse contra las calificaciones del supervisor después de seis meses en el trabajo. En la situación ideal, tales pruebas se validan durante periodos de ingreso abierto (o contratación abierta), de modo que es posible obtener el rango completo de resultados en el criterio. De esta manera, el empleo futuro de la prueba como instrumento de selección para excluir a los solicitantes con puntuaciones bajas dependerá de un fundamento sólido en datos de validación.

Cuando las pruebas se utilizan con propósitos de predicción, es necesario desarrollar una **ecuación de regresión**. Ésta describe la línea recta que se ajusta mejor para la estimación del criterio a partir de la prueba. No se analizará el enfoque estadístico para ajustar la línea recta, excepto para mencionar que reduce al mínimo la suma de desviaciones cuadradas a partir de la línea (Ghiselli, Campbell y Zedeck, 1981). Para los propósitos de esta obra, es más importante comprender la naturaleza y función de las ecuaciones de regresión.

Ghiselli, Campbell y Zedeck (1981) proporcionan un ejemplo sencillo de la regresión al servicio de la predicción, el cual se resume aquí. Supóngase que se intenta pronosticar el éxito en un empleo Y (valorado por el supervisor en una escala de siete puntos que va desde el desempeño deficiente hasta el excelente) a partir de las puntuaciones en una prueba de empleo X (con puntuaciones que van desde un nivel bajo de 0 hasta el límite superior de 100). La ecuación de regresión:

$$Y = .07X + .2$$

podría describir la línea recta que se ajusta mejor y, por tanto, producir las predicciones más precisas. Para un individuo que obtiene una puntuación de 55 en la prueba, el pronóstico del nivel de desempeño sería de 4.05 (es decir, .07(55) + .2). Una puntuación de 33 en la prueba produce un nivel de desempeño pronosticado de 2.51 (es decir, .07(33) + .2). Las predicciones adicionales se realizan de la misma manera.

COEFICIENTE DE VALIDEZ Y ERROR ESTÁNDAR DEL ESTIMADO

La relación entre las puntuaciones de una prueba y las medidas de criterio puede expresarse de varias maneras diferentes. Quizás el enfoque más popular consiste en calcular la correlación entre prueba y criterio (r_{xy}). En este contexto, la correlación resultante se conoce como **coeficiente de validez**. Entre más alto sea el coeficiente de validez r_{xy}, más precisa será la prueba en la predicción del criterio. En el caso hipotético donde r_{xy} fuera 1.00, la prueba poseería validez perfecta y daría lugar a predicciones impecables. Por supuesto, no existe una prueba así y los coeficientes de validez se encuentran de manera más habitual en los rangos de correlación de bajo a medio y en pocas ocasiones superan al .80. Pero ¿qué tan alto debería ser un coeficiente de validez? No existe una respuesta general para esta pregunta. Sin embargo, ésta puede abordarse de manera indirecta al investigar la relación entre el coeficiente de validez y el error del estimado correspondiente.

El **error estándar del estimado** (EE_{est}) es el margen de error que puede esperarse en la predicción de la puntuación de criterio, y se obtiene de la siguiente fórmula:

$$EE_{est} = DE_y \sqrt{1 - r_{xy}^2}$$

donde, r_{xy}^2 es el cuadrado del coeficiente de validez, y DE_y es la desviación estándar de las puntuaciones del criterio. Quizás el lector ha observado las semejanzas entre este índice y el error estándar de medición (EEM). De hecho, ambos índices ayudan a estimar los márgenes

de error. El *EEM* indica el margen de error de medición producido por la falta de confiabilidad de la prueba, mientras el EE_{est} indica el margen de error de predicción causado por la validez imperfecta de la prueba.

El EE_{est} ayuda a responder la pregunta fundamental: "¿con cuánta precisión puede predecirse el desempeño en el criterio a partir de las puntuaciones de una prueba?" (AERA, APA, NCME, 1985). Considérese la práctica común de tratar de pronosticar el promedio de calificaciones en la universidad a partir de las calificaciones de la preparatoria en una prueba de aptitud escolar. Para una prueba específica de aptitud, supóngase que se determina que el EE_{est} para pronóstico del promedio de calificaciones es de .2 (en la escala común de 0.0 a 4.0). ¿Qué significa esto para la persona evaluada para quien el pronóstico de calificaciones en la universidad es de 3.1? Como ocurre con todas las desviaciones estándar, el error estándar del estimado puede utilizarse para agrupar los resultados predichos en un sentido probabilístico. Suponiendo que la distribución de frecuencia de las calificaciones es normal, se sabe que las probabilidades son aproximadamente 68 entre 100 de que las calificaciones predichas para el examinando caigan entre 2.9 y 3.3 (más o menos un EE_{est}). De la misma manera, se sabe que las probabilidades son aproximadamente 95 entre 100 de que la calificación predicha de la persona caerá entre 2.7 y 3.5 (más o menos dos EE_{est}).

¿Cuál es la norma aceptable de exactitud en la predicción? No existe una respuesta sencilla para esta pregunta. Como podrá comprender el lector a partir de la siguiente discusión, las normas de exactitud en la predicción son, en parte, juicios de valor. Para explicar por qué ocurre esto, es necesario introducir los elementos básicos de la teoría de decisión (Taylor y Russell, 1939; Cronbach y Gleser, 1965).

APLICACIÓN DE LA TEORÍA DE DECISIÓN A LAS PRUEBAS PSICOLÓGICAS

Los defensores de la **teoría de decisión** destacan que el propósito de las pruebas psicológicas no es la medición en sí, sino la medición al servicio de la toma de decisiones. El administrador de personal desea saber a quién debe contratar; el funcionario de admisiones debe elegir a quién inscribir; la junta de libertad condicional desea conocer cuáles delincuentes constituyen un riesgo adecuado para preliberación y el psiquiatra necesita determinar cuáles pacientes requieren hospitalización.

El vínculo entre las pruebas y la toma de decisiones nunca es más obvio que en el contexto de los estudios de validez predictiva. Muchos de estos estudios utilizan resultados de prueba para determinar quién está en probabilidad de tener éxito o fracaso en la tarea criterio, de modo que, en el futuro, pueda excluirse a las personas con bajas puntuaciones en la prueba predictiva en el caso de inscripciones, empleo u otros privilegios. Ésta es la fundamentación por la que los funcionarios de admisiones o los contratantes requieren que los solicitantes obtengan una cierta puntuación mínima en un examen apropiado de ingreso o de empleo —pueden citarse estudios previos de validez predictiva para mostrar que los candidatos con calificaciones inferiores a un cierto punto de corte enfrentarán dificultades considerables en sus propósitos educativos o de empleo.

Con frecuencia, las pruebas psicológicas representan un papel de suma importancia en estos tipos de toma de decisiones institucionales. En una decisión institucional típica, un comité —o en ocasiones una sola persona— toma un gran número de decisiones comparables con base en una puntuación de corte en una o más pruebas de selección. A fin de presentar los conceptos clave de la teoría de decisión, permítasenos simplificar un poco las cosas y suponer que únicamente participa una sola prueba.

Aunque la mayor parte de las pruebas produce un rango de puntuaciones a lo largo de un continuo, en general es posible identificar una puntuación de corte o de aprobación/reprobación que divide la muestra en aquéllos para los cuales se pronostica el éxito contra aquéllos para los que se vaticina el fracaso en el criterio de interés. Supóngase que las personas con

predicción de éxito también se seleccionan para contratación o admisión. En este caso, la proporción de personas en el grupo de "pronóstico de éxito" se conoce como **proporción de selección**. Ésta puede variar de 0 a 1.0, dependiendo de la proporción de personas que se consideren como buenos candidatos para tener éxito en la medida de criterio.

Si los resultados de una prueba de selección permiten la dicotomía simple de "predicción de éxito" contra "predicción de fracaso", entonces el resultado subsecuente en la medida de criterio también puede dividirse de la misma manera en dos categorías, a saber, "tuvo éxito" y "fracasó". Desde esta perspectiva, todo estudio de validez predictiva produce una matriz de dos por dos, como se presenta en la figura 4–4.

Ciertas combinaciones de resultados predichos y reales son más probables que otras. Si una prueba tiene buena validez predictiva, entonces la mayoría de las personas para las que se pronostica el éxito lo tendrán y la mayoría de aquéllas para las que predice el fracaso, fracasarán; éstos son ejemplos de predicciones correctas y sirven para reforzar la validez de un instrumento de selección. Los resultados en estas dos casillas se conocen como **aciertos** porque la prueba ha hecho una predicción correcta.

DESEMPEÑO EN LA MEDIDA DE CRITERIO

PREDICCIÓN DE LA PRUEBA DE SELECCIÓN	Tuvo éxito	Fracasó
Tendrá éxito	Predicción correcta (acierto)	Falso positivo (error)
Fracasará	Falso negativo (error)	Predicción correcta (acierto)

Figura 4–4. Resultados posibles cuando se utiliza una prueba de selección para pronosticar el desempeño en una medida de criterio.

Pero ninguna prueba de selección es un medio perfecto de predicción, de modo que también son posibles otros tipos de resultados. Algunas personas para las que se predice el éxito fracasarán; éstos casos se conocen como **falsos positivos**; y algunas para las que se predice el fracaso, tendrán éxito si se les da oportunidad. Estos casos se conocen como **falsos negativos**. Los falsos positivos y falsos negativos se conocen, en general, como errores, porque en ambos casos la prueba ha hecho una predicción equivocada. Por último, la tasa de aciertos es la proporción de casos en los que la prueba pronostica de manera adecuada el éxito o el fracaso, es decir, tasa de acierto = (aciertos)/(aciertos + errores).

Los falsos positivos y falsos negativos son inevitables cuando se utilizan pruebas de selección en la vida real. La única manera de eliminar dichos errores de selección consistiría en desarrollar una prueba perfecta, un instrumento que tuviera un coeficiente de validez de +1.00, lo cual significa una correlación perfecta con la medida de criterio. Una prueba perfecta es posible en términos teóricos, pero nunca se ha observado que exista; no obstante, sigue siendo importante el desarrollo de pruebas de selección con validez predictiva muy elevada, a fin de reducir al mínimo los errores de decisión.

Los defensores de la teoría de decisión tienen dos suposiciones fundamentales acerca de la utilización de pruebas de selección:

1. El valor de diversos resultados para la institución puede expresarse en términos de una escala de utilidad común. Una escala de este tipo —pero de ninguna manera la única— es la de pérdidas y ganancias. Por ejemplo, cuando se utiliza un inventario de intereses para seleccionar a vendedores, una empresa puede anticipar ganancias de los solicitantes que se identifiquen de manera correcta como exitosos, pero perderá dinero cuando, de manera inevitable, algunos de los seleccionados no vendan lo suficiente ni siquiera para costear su propio salario (falsos positivos). El costo del procedimiento de selección

también debe tomarse en cuenta dentro de la escala de utilidad.

2. En las decisiones de selección en instituciones, la estrategia que en general es más útil es aquella que aumenta al máximo la ganancia promedio en la escala de utilidad (o reduce al mínimo la pérdida promedio) sobre muchas decisiones similares. Por ejemplo, ¿cuál proporción de selección produce la ganancia promedio más grande en la escala de utilidad? Por tanto, la maximización es el principio fundamental de la decisión.

La aplicación de la teoría de decisión es mucho más complicada de lo que se ilustra aquí, principalmente debido a la dificultad para encontrar una escala de utilidad común para diferentes resultados. Considérese la difícil situación del funcionario de admisiones de cualquier universidad grande. Si la proporción de selección es bastante estricta, entonces la mayoría de los estudiantes admitidos tendrá éxito, pero algunos no admitidos también podrían haber tenido éxito y, por tanto, se perdería su apoyo económico a la universidad (matrícula, cuotas). No obstante, si la proporción de selección es demasiado indulgente, entonces el porcentaje de falsos positivos (estudiantes admitidos que tendrán un fracaso subsecuente) se dispara. ¿Cómo ha de calcularse el costo de un falso positivo? Se puede calcular el costo económico —por ejemplo, los asesores dedican un cierto número de horas con un salario especificado a orientar a estos alumnos. Pero ninguna escala de utilidad puede englobar las otras consecuencias diversas, como la necesidad de servicios adicionales de regularización (que requieren dinero), el aumento en el cinismo de los maestros (una cuestión de moral del personal) y las esperanzas perdidas de estudiantes engañados (¡cuya decepción afecta la percepción del público acerca de la universidad e incluso pueden influir en los subsidios gubernamentales futuros!). Es claro que los conceptos estadísticos precisos de la teoría de decisión simplifican en extremo las influencias complejas que determinan la utilidad en el mundo real.

Sin embargo, en ambientes institucionales grandes, donde puede identificarse una escala de utilidad común, los principios de la teoría de decisión se pueden aplicar a los problemas de selección con resultados que instan a la reflexión. Por ejemplo, Schmidt, Hunter, McKenzie y Muldrow (1979) analizaron el impacto potencial de utilizar la *Programmer Aptitude Test* (Prueba de Aptitud del Programador; PAT, Hughes y McNamara, 1959) en la selección de programadores de computadoras para el gobierno federal. Basaron su análisis en los siguientes hechos y suposiciones:

1. Las puntuaciones de la PAT y las medidas posteriores de desempeño en programación dentro del trabajo se correlacionan de manera bastante sustancial; el coeficiente de validez de la PAT es .76 (hecho).
2. El gobierno contrata a 600 nuevos programadores cada año (hecho).
3. El costo de la prueba es de cerca de 10 dólares por persona examinada (hecho).
4. Los programadores permanecen en el empleo aproximadamente nueve años y reciben aumentos de sueldo de acuerdo con una tasa de salarios específica (hecho).
5. Los supervisores pueden estimar de manera precisa la productividad anual en dólares de los programadores con desempeño bajo, promedio y superior (suposición).

Con base en estos hechos y suposiciones, Schmidt y colaboradores (1979) compararon la utilización hipotética de la PAT contra otros procedimientos de selección de menor validez. Dado que la utilidad de una prueba se determina en parte por el porcentaje de solicitantes que se seleccionan para el empleo, los investigadores también examinaron el impacto de diferentes proporciones de selección sobre la productividad total. En cada caso, estimaron el aumento anual en productividad, valuada en dólares, de utilizar la PAT en vez de un procedimiento alternativo y menos eficaz. En general, se estimó que el uso de la PAT aumentaba la productividad en decenas de millones de dólares. El aumento específico estimado dependió de la proporción de selección y del coeficiente de validez de los procedimientos alternativos

hipotéticos (cuadro 4–1). Por ejemplo, si se contrataba a 80% de los solicitantes (proporción de selección de .80), la utilización de la PAT aumentaría la productividad del gobierno federal en, cuando menos, 5.6 millones de dólares (si el procedimiento alternativo tuviera un coeficiente de validez de .50) y posiblemente hasta 16.5 millones de dólares (si dicho procedimiento no tuviera alguna validez). Si la proporción de selección fuera bastante pequeña, el uso de la PAT para la selección elevaría la productividad aún más —posiblemente hasta cerca de 100 millones de dólares. Schmidt y colaboradores (1979) concluyeron que "el impacto de los procedimientos válidos de selección en la productividad de la fuerza laboral es considerablemente más grande de lo que ha creído la mayoría de los psicólogos dedicados a la selección de personal".

En general, las grandes organizaciones que toman docenas o cientos de decisiones similares de contratación cada año, pueden beneficiarse de la aplicación de la teoría de decisión a los procedimientos de selección, mientras que es posible que las empresas más pequeñas no posean una base de datos suficiente para aplicar las metodologías complejas inherentes a este enfoque. Cuando es posible, la utilización del principio de maximización en conjunto con una escala mensurable de utilidad, esta teoría es muy prometedora. Por ejemplo, Schmidt, Hunter, Outerbridge y Trattner (1986) estiman que el uso de medidas válidas de capacidad cognitiva en lugar de procedimientos ajenos a pruebas (p. ej., requerir un título específico) para la selección de personal, aumentaría la productividad de los trabajadores del gobierno federal en 8 mil millones de dólares en el curso del cargo laboral típico de 13 años.

Cuadro 4–1. Incremento de la productividad estimada (en millones de dólares) para la utilización a un año de la Prueba de aptitud del programador

Proporción de selección	Validez verdadera del procedimiento				
	.00	.20	.30	.40	.50
.05	97.2	71.7	58.9	46.1	33.3
.10	82.8	60.1	50.1	39.2	28.3
.20	66.0	48.6	40.0	31.3	22.6
.30	54.7	40.3	33.1	25.9	18.7
.40	45.6	34.6	27.6	21.6	15.6
.50	37.6	27.7	22.8	17.8	12.9
.60	30.4	22.4	18.4	14.4	10.4
.70	23.4	17.2	14.1	11.1	8.0
.80	16.5	12.2	10.0	7.8	5.6

Fuente: Reimpreso con autorización de Schmidt, F. L., Hunter, J. E., McKenzie, R. C. y Muldrow, T. W. (1979). Impact of valid selection procedures on work-force productivity. *Journal of Applied Psychology, 64*, 609-626.

TABLAS TAYLOR-RUSSELL

En el contexto de la teoría de decisión, deberían describirse las tablas estadísticas publicadas por Taylor y Russell (1939), las cuales permiten que el usuario de pruebas determine la proporción esperada de solicitantes exitosos seleccionados con la utilización de una de ellas. Por obvias razones, estas pautas se conocen como tablas Taylor-Russell. A fin de utilizarlas, el examinador debe especificar: 1) la validez predictiva de la prueba, 2) la proporción de selección y 3) la tasa base para los solicitantes exitosos. Un cambio en cualesquiera de estos factores alterará la precisión de la prueba para la selección.

En general, la validez predictiva de una prueba se conoce a partir de estudios anteriores y consiste en r_{xy}, la correlación entre la prueba y el criterio. La proporción de selección es la cantidad de solicitantes que se seleccionan (en general debido a que se pronostica que tendrán éxito). La **tasa base** es la proporción de solicitantes exitosos que se seleccionarían con los métodos actuales, sin el beneficio de la nueva prueba. (En el caso extremo, la tasa base es la proporción de solicitantes exitosos que se elegirían de manera aleatoria sin el beneficio de algún procedimiento de selección.) Cuando se conocen estos tres factores, se pueden consultar las tablas Taylor-Russell para determinar

la proporción de éxitos esperados a través de la aplicación de la prueba. De esta manera, el usuario de ésta puede determinar el grado al cual el empleo de una nueva mejoraría la selección por encima de la tasa base obtenida con los métodos existentes.

Quizás un ejemplo específico ayudará a aclarar la aplicación potencial de las tablas Taylor-Russell (cuadro 4–2). Supóngase que la tasa base para los solicitantes exitosos es de .60, lo cual significa que 60% de los solicitantes aceptados con los métodos actuales resulta ser exitoso. Supóngase también que la proporción de selección es .50, lo cual significa que se selecciona a 50% de todos los solicitantes. Lo que es más, imagínese que una nueva prueba tiene un coeficiente de validez de .40, que especifica la correlación entre puntuaciones de prueba y criterio. Bajo estas suposiciones, las tablas Taylor-Russell suministran la proporción esperada de éxitos mediante la utilización de la nueva prueba, que resulta ser de .73, una mejoría sustancial por encima de la tasa base actual de .60 para la selección exitosa.

La conclusión más fascinante que surge de las tablas Taylor-Russell es que las pruebas con validez "deficiente" pueden, de todas maneras, mejorar a un nivel sustancial la precisión en la selección —si la proporción de selección es suficientemente baja. Considérese una prueba con validez de tan sólo .20, que no parece muy impresionante; por lo común, en la mayoría de las pruebas la mayor parte de los coeficientes de validez son mucho más altos. Supóngase también

Cuadro 4–2. Proporción esperada de éxitos para una prueba de selección de una validez dada, con una proporción específica de selección, para una tasa base de .60

	Proporción de selección										
Validez	**.05**	**.10**	**.20**	**.30**	**.40**	**.50**	**.60**	**.70**	**.80**	**.90**	**.95**
.00	.60	.60	.60	.60	.60	.60	.60	.60	.60	.60	.60
.05	.64	.63	.63	.62	.62	.62	.61	.61	.61	.60	.60
.10	.68	.67	.65	.64	.64	.63	.63	.62	.61	.61	.60
.15	.71	.70	.68	.67	.66	.65	.64	.63	.62	.61	.61
.20	.75	.73	.71	.69	.67	.66	.65	.64	.63	.62	.61
.25	.78	.76	.73	.71	.69	.68	.66	.65	.63	.62	.61
.30	.82	.79	.76	.73	.71	.69	.68	.66	.64	.62	.61
.35	.85	.82	.78	.75	.73	.71	.69	.67	.65	.63	.62
.40	.88	.85	.81	.78	.75	.73	.70	.68	.66	.63	.62
.45	.90	.87	.83	.80	.77	.74	.72	.69	.66	.64	.62
.50	.93	.90	.86	.82	.79	.76	.73	.70	.67	.64	.62
.55	.95	.92	.88	.84	.81	.78	.75	.71	.68	.64	.62
.60	.96	.94	.90	.87	.83	.80	.76	.73	.69	.65	.63
.65	.98	.96	.92	.89	.85	.82	.78	.74	.70	.65	.63
.70	.99	.97	.94	.91	.87	.84	.80	.75	.71	.66	.63
.75	.99	.99	.96	.93	.90	.86	.81	.77	.71	.66	.63
.80	1.00	.99	.98	.95	.92	.88	.83	.78	.72	.66	.63
.85	1.00	1.00	.99	.97	.95	.91	.86	.80	.73	.66	.63
.90	1.00	1.00	1.00	.99	.97	.94	.88	.82	.74	.67	.63
.95	1.00	1.00	1.00	1.00	.99	.97	.92	.84	.75	.67	.63
1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.86	.75	.67	.63

Fuente: Reproducido de Taylor, H. C. y Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection. *Journal of Applied Psychology, 23*, 565-578. Dominio público.

que la tasa base para la selección exitosa es de .60, lo cual probablemente es una tasa base realista para muchas formas de selección de personal. Lo que es más, supóngase que la proporción de selección es muy rigurosa, digamos .05, lo cual significa que sólo se considera aceptable a 5% de los solicitantes y, por tanto, es a quienes se selecciona. Bajo estas suposiciones, la proporción esperada de éxitos a través del uso de la prueba es de .75, una mejoría neta de 15% por encima de la tasa base de .60.

VALIDEZ DE CONSTRUCTO

El tipo final de validez que se analiza en esta unidad es la validez de constructo, la cual es, sin duda, la más difícil y esquiva de todas. Un **constructo** es una cualidad o rasgo teórico, intangible, en el cual difieren los individuos (Messick, 1995). Ejemplos de constructos incluyen capacidad de liderazgo, hostilidad sobrecontrolada, depresión e inteligencia. Obsérvese en cada uno de estos ejemplos que los constructos se infieren de la conducta, pero son más que ésta en sí. En general, a un nivel teórico se considera que los constructos tienen cierta forma de existencia independiente y ejercen influencias amplias, pero en cierto grado posibles de predecir, sobre la conducta humana. Una prueba que se diseñe para medir un constructo debe estimar la existencia de una característica inferida, subyacente (p. ej., capacidad de liderazgo), con base en una muestra limitada de conducta. La validez de constructo se refiere a la idoneidad de estas inferencias acerca del constructo subyacente.

Todos los constructos psicológicos poseen dos características en común:

1. No existe un solo referente externo suficiente para validar la existencia del constructo, es decir, éste no puede definirse en términos operacionales (Cronbach y Meehl, 1955).
2. No obstante, puede obtenerse una red de suposiciones entrelazadas a partir de la teoría existente acerca del constructo (AERA, APA, NCME, 1985).

Se ilustrarán estos elementos con referencia al constructo de la psicopatía (Cleckley, 1976), una constelación de personalidad que se caracteriza por conducta antisocial (mentira, robo y violencia ocasional), una falta de culpa y vergüenza e impulsividad.* Con toda seguridad, la psicopatía es un constructo en cuanto a que no existe una sola característica o resultado conductual suficiente para determinar quién presenta una fuerte psicopatía y quién no. En promedio, podría esperarse que a los psicópatas se les encarcelara con frecuencia, pero eso también ocurre con la mayoría de los delincuentes comunes. Lo que es más, de alguna manera muchos psicópatas exitosos evitan que se les aprehenda alguna vez (Cleckley, 1976). La psicopatía no puede estimarse sólo a partir de los roces con la ley.

Sin embargo, a partir de la teoría existente sobre la psicopatía se puede deducir una red de suposiciones entrelazadas. Se supone que el problema fundamental en la psicopatía es una deficiencia en la capacidad para sentir excitación emocional —ya sea empatía, culpa, temor al castigo o ansiedad bajo estrés (Cleckley, 1976). De esta valoración se desprenden varias predicciones. Por ejemplo, los psicópatas deben mentir de manera convincente, tener mayor tolerancia al dolor físico, mostrar menor excitación autónoma en estado de reposo y meterse en problemas debido a su falta de inhibición conductual. Así, para validar una medida de la psicopatía, necesitarían verificarse varias diferentes expectativas basadas en esta teoría acerca de ella.

La **validez de constructo** atañe a las pruebas psicológicas que afirman medir atributos psicológicos complejos, multifacéticos y vinculados con la teoría como la psicopatía, inteligencia, capacidad de liderazgo y similares. El punto crucial a comprender acerca de la validez de constructo es que "ningún criterio o universo de contenido se acepta como completamente adecuado para definir la cualidad que se mide" (Cronbach y Meehl, 1955). Por tanto, la demostración de la validez de constructo siempre

* El constructo de la psicopatía es muy similar a lo que ahora se denomina trastorno de personalidad antisocial (American Psychiatric Association, 1994).

depende de un programa de investigación que utilice los diversos procedimientos que se esquematizan en las siguientes secciones. Para evaluar la validez de constructo de una prueba, debe acumularse una variedad de evidencia a partir de numerosas fuentes.

Aunque establecer la validez de constructo de una prueba es un proceso largo y complejo, los diversos procedimientos están diseñados para responder a una pregunta crucial: con base en los conceptos teóricos actuales sobre el constructo que la prueba afirma medir, ¿se encuentran los tipos de relaciones con criterios ajenos a la prueba que la teoría predice? Considérese el concepto de psicopatía, que se analizó antes, como lo mide la escala de Desviación psicopática (Dp) del MMPI y el MMPI-2. Un pequeño trozo de evidencia que apoya la validez de constructo de esta escala es el descubrimiento de que los cazadores que "por descuido" le habían disparado a una persona tenían elevaciones significativas en Dp en comparación con otros cazadores (Cronbach y Meehl, 1955). Dicha conclusión encaja de manera adecuada con el concepto teórico de psicopatía, en especial en lo que se refiere a la falta de inhibición conductual. Por supuesto, se necesitarían muchas otras líneas de evidencia para confirmar la validez de constructo de la escala Dp. Por tanto, puede observarse que la investigación de la validez de constructo no es esencialmente diferente de los procedimientos científicos generales para confirmar una teoría.

Muchos teóricos en psicometría consideran a la validez de constructo el concepto unificador para todos los tipos de evidencia de validez (Cronbach, 1988; Guion, 1980; Messick, 1995). Según este punto de vista, los estudios individuales de validez de contenido, concurrente y predictiva se consideran tan sólo como evidencia de apoyo en la búsqueda acumulativa de validación del constructo.

ENFOQUES DE LA VALIDEZ DE CONSTRUCTO

¿Cómo es que un autor de pruebas determina si un nuevo instrumento posee validez de constructo? Como se indicó antes, ningún procedimiento único será suficiente para esta difícil tarea. La evidencia de la validez de constructo puede encontrarse en casi cualquier estudio empírico que examine las puntuaciones de prueba de grupos apropiados de individuos. La mayor parte de los estudios de validez de constructo caen dentro de una de las siguientes categorías:

- Análisis para determinar si las subpruebas o reactivos de una prueba son homogéneos y, por tanto, miden un solo constructo.
- Estudio de los cambios relacionados con el desarrollo para determinar si son consistentes con la teoría del constructo.
- Investigación para evaluar si las diferencias grupales en las puntuaciones de una prueba son consistentes con la teoría.
- Análisis para determinar si los efectos de la intervención sobre las puntuaciones de una prueba son consistentes con la teoría.
- Correlación de la prueba con otras pruebas y medidas relacionadas y no relacionadas.
- Análisis factoriales de las puntuaciones de prueba en relación con otras fuentes de información.

A continuación se analizan con mayor detalle estas fuentes de evidencia de validez de constructo.

HOMOGENEIDAD DE UNA PRUEBA

Si una prueba mide un solo constructo, entonces es probable que los reactivos (o subpruebas) que la componen sean homogéneos (esto también se conoce como **consistencia interna**). En la mayor parte de los casos, la homogeneidad se incorpora dentro de la prueba durante el proceso de desarrollo que se analiza con mayor detalle en la siguiente unidad. La finalidad del desarrollo de una prueba es la selección de reactivos que formen una **escala homogénea**. El método que más se utiliza para lograr esta meta es la correlación de cada reactivo potencial con la puntuación total y la selección de

reactivos que muestren altas correlaciones con la puntuación total. Un procedimiento análogo consiste en correlacionar las subpruebas con la puntuación total en las primeras fases del desarrollo de la prueba. De este modo, pueden revisarse las escalas irregulares que no se correlacionan en algún grado mínimo con la puntuación total antes de que el instrumento se publique para utilización general.

La homogeneidad es un primer paso importante en la certificación de la validez de constructo de una prueba nueva, pero en sí misma, es una evidencia débil. Kline (1986) ha resaltado circularidad del procedimiento:

> Si todos los reactivos en una reserva de éstos estuvieran muy lejos del blanco y no midieran aquello que esperamos, se seleccionarían según el criterio de su correlación con la puntuación total, lo cual nunca puede funcionar. Ha de señalarse que el mismo argumento se aplica a la factorización de la reserva de reactivos. Incluso es posible que los reactivos deficientes tengan un factor general. Esta objeción es sensata y se ha de refutar de manera empírica. Aunque se haya encontrado un conjunto de reactivos homogéneos a través de un análisis de reactivos, de todas maneras tiene que presentarse evidencia acerca de su validez. Así, construir una prueba homogénea no es suficiente, se deben llevar a cabo estudios de validez.

Además de demostrar la homogeneidad de los reactivos, un autor de una prueba debe proporcionar otras múltiples fuentes de validez de constructo, las cuales se analizan a continuación.

CAMBIOS APROPIADOS RELACIONADOS CON EL DESARROLLO

Puede suponerse que muchos constructos muestran cambios regulares relacionados con la edad, desde la infancia temprana hasta la edad adulta y quizá más allá. Considérese como ejemplo el constructo del conocimiento de vocabulario. Desde los inicios de las pruebas de inteligencia, a principios del siglo XX, se sabe que el conocimiento del vocabulario aumenta de manera exponencial desde la infancia temprana hasta la infancia tardía. Investigaciones más recientes demuestran que el vocabulario aumenta, aunque a un paso más lento, hasta la vejez (Gregory y Gernert, 1990). Entonces, para cualquier prueba nueva de vocabulario, un trozo importante de evidencia de validez de constructo sería que los individuos de mayor edad tienen mejores puntuaciones que aquéllos más jóvenes, suponiendo que se mantengan constantes los factores de educación y salud.

Por supuesto, no todos los constructos se prestan a pronósticos acerca de cambios vinculados con el desarrollo. Por ejemplo, no es claro si una escala que mida "asertividad" debería mostrar un patrón de aumento, disminución o estabilidad de las puntuaciones a medida que se avanza en edad. Los cambios del desarrollo serían irrelevantes para la validez de constructo de una escala de este tipo. También se debería mencionar que los cambios apropiados del desarrollo constituyen solamente una pieza del rompecabezas de la validez de constructo. Este enfoque no proporciona información acerca de cuál es la relación del constructo con otros.

DIFERENCIAS GRUPALES CONSISTENTES CON LA TEORÍA

Una manera de reforzar la validez de un nuevo instrumento consiste en mostrar que, en promedio, las personas con antecedentes y características diferentes obtienen puntuaciones de prueba consistentes con la teoría. En términos específicos, las personas que se consideran altas en el constructo que mide la prueba deberían obtener puntuaciones altas, mientras las personas con cantidades supuestamente bajas del constructo deberían obtener puntuaciones bajas.

Crandall (1981) elaboró una escala de interés social que ilustra el empleo de las diferencias grupales consistentes con la teoría en el proceso de validación del constructo. Con base en la teoría de Alfred Adler, Crandall (1984) definió el **interés social** como un "interés y preocupación acerca de los demás". Para me-

dir este constructo, diseñó un instrumento breve y sencillo que consta de 15 reactivos de elección forzada. En cada uno, una de las dos alternativas incluye un rasgo que se relaciona de manera estrecha con el concepto de interés social (p. ej., servicial), mientras la otra opción consiste en un rasgo igualmente atractivo, pero sin relación con lo social (p. ej., viveza de ingenio). Al individuo se le instruye para que "elija el rasgo que usted valore más". Cada uno de los 15 reactivos se califican con uno si se elige el rasgo de interés social y con cero si se elige el otro; así, las puntuaciones totales en la *Social Interest Scale* (Escala de Interés Social; SIS) pueden ir de 0 a 15.

El cuadro 4–3 presenta las puntuaciones promedio de 13 grupos bien definidos de participantes. El lector observará que los individuos en probabilidad de tener un alto nivel de interés social (p. ej., monjas) obtienen las puntuaciones promedio más altas en la SIS, mientras las más bajas las obtienen personas supuestamente egocéntricas (p. ej., modelos) y aquéllos con evidente conducta antisocial (delincuentes). Estas conclusiones son consistentes con la teoría y sustentan la validez de constructo de este interesante instrumento.

EFECTOS DE LA INTERVENCIÓN CONSISTENTES CON LA TEORÍA

Otro enfoque de la validez de constructo consiste en mostrar que las puntuaciones de una prueba cambian en la dirección y cantidad apropiadas en reacción a intervenciones planeadas o no planeadas. Por ejemplo, las puntuaciones de los ancianos en una batería de prueba de orientación espacial deberían de aumentar después de que éstos recibieran entrenamiento cognitivo diseñado de manera específica para aumentar sus capacidades de orientación espacial. En términos más precisos, si la batería de una prueba posee validez de constructo, se puede predecir que las puntuaciones en orientación espacial deberían mostrar mayor incremento de la preprueba a la posprueba que el encontrado en capacidades no relacionadas que no son el objetivo del entrenamiento especial (p. ej., razonamiento inductivo, velocidad perceptual, razonamiento numérico o razonamiento verbal). Willis y Schaie (1986) encontraron un patrón de este tipo en resultados de prueba en un estudio de entrenamiento cognitivo con personas ancianas, lo cual sustenta la validez de constructo de su medida de orientación espacial.

Cuadro 4–3. Puntuaciones medias en la Escala de interés social para grupos seleccionados

Grupo	N	Puntuación media
Hermanas ursulinas	6	13.3
Miembros adultos de la iglesia	147	11.2
Voluntarios para caridad	9	10.8
Estudiantes de educación media distinguidos por alto interés social	23	10.2
Estudiantes universitarios distinguidos por alto interés social	21	9.5
Empleados universitarios	327	8.9
Estudiantes universitarios	1 784	8.2
Estudiantes universitarios distinguidos por bajo interés social	35	7.4
Modelos profesionales	54	7.1
Estudiantes de educación media distinguidos por bajo interés social	22	6.9
Ateos y agnósticos adultos	30	6.7
Delincuentes convictos	30	6.4

Fuente: Adaptado con autorización de Crandall, J. (1981). *Theory and measurement of social interest: Empirical tests of Alfred Adler's concept.* Nueva York: Columbia University Press.

VALIDEZ CONVERGENTE Y DE DISCRIMINACIÓN

La **validez convergente** se demuestra cuando una prueba se correlaciona mucho con otras variables o pruebas con las que comparte una superposición de constructos. Por ejemplo, dos pruebas diseñadas para medir diferentes tipos de inteligencia deberían compartir de todas maneras una cantidad considerable del factor general en inteligencia como para producir una fuerte correlación (digamos .5 o más) cuando se aplican de manera conjunta a una muestra heterogénea de individuos. De hecho, cualquier prueba nueva de inteligencia que no se correlacione al menos en cierto grado con las medidas existentes sería sumamente sospechosa, con base en que no posee validez convergente.

La **validez de discriminación** se demuestra cuando una prueba no se correlaciona con variables o pruebas con las cuales debería tener diferencias. Por ejemplo, el interés social y la inteligencia no tienen relación teórica, y las pruebas con estos dos constructos deberían correlacionarse en sentido negativo, si es que acaso lo hacen.

En un artículo clásico que con frecuencia se cita, pero en pocas ocasiones se emula, Campbell y Fiske (1959) propusieron un diseño experimental sistemático para confirmar de manera simultánea la validez convergente y la de discriminación de una prueba psicológica. Su diseño se denomina **matriz multirrasgo-multimétodo** y requiere de la evaluación de dos o más rasgos a través de dos o más métodos. El cuadro 4–4 proporciona un ejemplo hipotético de este enfoque. En este ejemplo, se miden tres rasgos (**A, B** y **C**) a través de tres métodos (1, 2 y 3). Por ejemplo, los rasgos **A, B** y **C** podrían ser interés social, creatividad y dominio. Los métodos 1, 2 y 3 podrían ser un inventario de autoinforme, calificaciones de compañeros y una prueba proyectiva. Así, **A1** representaría un inventario de autoinforme de interés social, **B2** una calificación de compañeros acerca de creatividad, **C3** una medida de dominio obtenida de una prueba proyectiva y así sucesivamente.

Obsérvese que en este ejemplo se estudiaron nueve pruebas (los tres rasgos se miden, cada uno, con tres métodos). Cuando cada una de éstas se aplica en dos ocasiones al mismo grupo de individuos y se correlacionan las puntuaciones en todos los pares de pruebas, el resultado es una **matriz multirrasgo-multimétodo** (cuadro 4–4). Esta matriz es una fuente

Cuadro 4–4. Matriz hipotética multirrasgo-multimétodo

		Método 1			Método 2			Método 3		
	Rasgos	A_1	B_1	C_1	A_2	B_2	C_2	A_3	B_3	C_3
Método 1	A_1	(89)								
	B_1	51	(89)							
	C_1	38	37	(76)						
Método 2	A_2	57	22	09	(93)					
	B_2	22	57	10	68	(94)				
	C_2	11	11	46	59	58	(84)			
Método 3	A_3	56	22	11	67	42	33	(94)		
	B_3	23	58	12	43	66	34	67	(92)	
	C_3	11	11	45	34	32	58	58	60	(85)

Nota: Las letras *A, B* y *C* se refieren a los rasgos; los subíndices 1, 2 y 3 se refieren a los métodos. La matriz consiste de coeficientes de correlación (se omitieron los decimales). Véase texto.

Fuente: Reproducido con autorización de Campbell, D. T. y Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

prolífica de datos sobre confiabilidad, validez convergente y validez de discriminación.

Por ejemplo, las correlaciones en la diagonal principal (entre paréntesis) son coeficientes de confiabilidad para cada prueba. Mientras más altos sean estos valores, mejor y de preferencia sería mejor que aquí se observaran valores en los rangos de .80 o .90. Las correlaciones en las tres diagonales más pequeñas (en negritas) proporcionan evidencia sobre la validez convergente —medición del mismo rasgo a través de diferentes métodos. Estas correlaciones deben ser fuertes y positivas, como se muestra aquí. Véase que el cuadro también incluye correlaciones entre los diferentes rasgos que se miden con el mismo método (en los triángulos sólidos) y diferentes rasgos medidos con diferentes métodos (triángulos punteados). Estas correlaciones deben ser las más bajas de todas dentro de la matriz, ya que proporcionan evidencia de validez discriminativa.

La metodología de Campbell y Fiske (1959) es una contribución importante para la comprensión del proceso de validación de una prueba. Sin embargo, poner en práctica de manera completa este procedimiento requiere por lo común de un compromiso monumental por parte de los investigadores. Es más frecuente que los autores de pruebas obtengan datos sobre validez convergente y discriminativa en retazos, en lugar de producir una matriz completa de intercorrelaciones. Meier (1984) proporciona una de las pocas ejecuciones realistas de la matriz multirrasgo-multimétodo en un examen de la validez del constructo de "agotamiento".

ANÁLISIS FACTORIAL

El análisis factorial es una técnica estadística especializada que es de particular utilidad para la investigación de la validez de constructo. En el tema 8A, Pruebas de aptitud y análisis factorial, se analiza con mayores detalles esta técnica; aquí se proporciona un avance rápido, de modo que el lector pueda apreciar el papel del análisis factorial en el estudio de la validez de constructo. El propósito del **análisis factorial** consiste en identificar el número mínimo de determinantes (factores) que se requieren para explicar las intercorrelaciones entre una batería de pruebas. El objetivo en el análisis factorial es encontrar el conjunto más pequeño de dimensiones, llamadas **factores**, que puedan explicar la disposición observada de intercorrelaciones entre las pruebas individuales. Un enfoque típico en el análisis factorial consiste en aplicar una batería de pruebas a varios cientos de personas y después calcular una matriz de correlación a partir de las puntuaciones en todos los posibles pares de pruebas. Por ejemplo, si se aplican 15 pruebas a una muestra de pacientes psiquiátricos y neurológicos, el primer paso en el análisis factorial consiste en calcular las correlaciones entre las puntuaciones de los 105 pares posibles de pruebas.* Aunque puede ser factible observar ciertos agrupamientos de pruebas que midan rasgos comunes, es más típico que la mayor cantidad de datos que se encuentran en una matriz de correlación sea simplemente demasiado compleja como para el que el ojo humano la analice de manera eficaz sin ninguna ayuda. Por fortuna, los procedimientos de análisis factorial ejecutados por computadora buscan este patrón de intercorrelaciones, identifican un pequeño número de factores y después producen una tabla de cargas factoriales. De hecho, una **carga factorial** es una correlación entre una prueba individual y un solo factor. Así, las cargas factoriales pueden variar entre –1.0 y +1.0. El resultado final del análisis factorial es una tabla que presenta la correlación de cada prueba con cada factor.

Una tabla de cargas factoriales ayuda a describir la composición factorial de una prueba y, por tanto, proporciona información relacionada con la validez de constructo. Se ilustrará este punto con los datos analítico-factoriales de un estudio de la Prueba de categorías. Esta prueba es una relativamente compleja de formación

* La fórmula general para el número de apareamientos entre N pruebas es $N(N-1)/2$. Así, se aplican 15 pruebas, habrá $15 \times 14/2$ o 105 posibles pares de pruebas individuales.

de conceptos diseñada para ser diferente de las medidas psicométricas tradicionales de la inteligencia y superior a ellas en la detección de trastornos neurológicos (Reitan y Wolfson, 1985). Si en realidad la Prueba de categorías mide algo diferente de las pruebas tradicionales de inteligencia, entonces debería tener fuertes cargas en uno o más factores no representados por las subpruebas del WAIS. Dicha conclusión fortalecería la validez de constructo de la Prueba de categorías al distinguirla de las medidas tradicionales de inteligencia.

Lansdell y Donnelly (1977) aplicaron la Escala Wechsler de Inteligencia para Adultos, la Prueba de categorías y la Prueba de golpeteo de dedos a 94 pacientes psiquiátricos y neurológicos. Las puntuaciones de prueba se sometieron a un análisis factorial que produjo las cargas factoriales que se muestran en el cuadro 4–5. Obsérvese que las subpruebas verbales del WAIS tienen las cargas más elevadas en el factor I, que seguramente es un factor de comprensión verbal. La Prueba de categorías tienen una carga mínima en este factor, lo cual indica que las capacidades verbales no son de particular importancia para el desempeño adecuado en esta prueba. Diseño con cubos (.74) y Ensamble de objetos (.73) tienen sus cargas más fuertes en el factor II y, por lo común, se les coloca en el factor de organización perceptual.* Por desgracia, la Prueba de categorías tiene una carga sustancial (.82) en este factor y sólo en éste. Al menos para esta muestra de participantes, parece ser que la Prueba de categorías es tan sólo una medida alternativa de las habilidades de organización perceptual y no es una prueba nueva y diferente, como muchos de sus usuarios querrían afirmar. De manera incidental, debe decirse que el factor III parece medir libertad de la distractibilidad y el factor IV parece ser una medida pura de velocidad motora.

* Obsérvese que los seres humanos son quienes proporcionan la etiqueta para un factor, con base en un análisis de las pruebas que tienen las cargas más elevadas en él. Es concebible que dos investigadores pudieran utilizar nombres diferentes para el mismo factor —por ejemplo, referirse al factor II como **organización perceptual** como **análisis visoespacial**.

Cuadro 4–5. Cargas factoriales para la Prueba de categorías, Prueba de golpeteo con los dedos y subpruebas del WAIS

	Carga factorial			
Prueba	**I**	**II**	**III**	**IV**
Información	.88	.15	.07	.07
Comprensión	.83	−.03	.06	−.09
Aritmética	.43	.26	.67	−.12
Semejanzas	.78	.30	.17	.02
Retención de dígitos	.23	.08	.83	.12
Vocabulario	.92	.07	.06	.01
Claves	.25	.31	.21	.61
Figuras incompletas	.64	.50	−.24	−.01
Diseño con cubos	.39	.74	.06	.20
Ordenación de dibujos	.50	.60	.12	−.01
Composición de objetos	.29	.73	.00	.31
Categorías	.19	.82	.11	−.18
Prueba de golpeteo con los dedos	.07	−.08	.18	.76

Fuente: Adaptado con autorización de Lansdell, H. y Donnelly, E. F. (1977). Factor analysis of the Wechsler Adult Intelligence Scale Subtests and the Halstead-Reitan Category and Tapping Tests. *Journal of Consulting and Clinical Psychology, 45*, 412-416.

PREOCUPACIONES EXTERNAS A LA VALIDEZ Y AMPLIACIÓN DEL ESPECTRO DE LA VALIDEZ DE UNA PRUEBA

Se comienza esta sección con una revisión de las **preocupaciones externas a la validez**, que incluyen los efectos colaterales y consecuencias involuntarias de las pruebas. Al reconocer la importancia de los aspectos externos a la validez, los psicólogos confirman que la decisión de utilizar una prueba implica consideraciones sociales, legales y políticas que van más allá de las cuestiones tradicionales de la validez técnica. En una explicación relacionada, también se revisará la manera en que el interés en las preocupaciones externas a la validez ha

estimulado a varios teóricos a ampliar este concepto. Como descubrirá el lector, las implicaciones de valor y las consecuencias sociales están incluidas ahora dentro del espectro cada vez más amplio de la validez de una prueba.

Incluso si una prueba es válida, imparcial y justa, la decisión de utilizarla puede estar gobernada por consideraciones adicionales. Cole y Moss (1989) plantean los siguientes factores:

- ¿Cuál es el propósito para el que se emplea la prueba?
- ¿A qué grado se logran los propósitos a través de las acciones que se toman?
- ¿Cuáles son los posibles efectos colaterales o consecuencias involuntarias del empleo de una prueba?
- ¿Cuáles alternativas posibles para una prueba podrían satisfacer el mismo propósito?

Aquí sólo se revisan las preocupaciones externas a la validez que tienen mayor relevancia y se muestra cómo han servido para ampliar el espectro de la validez de una prueba. Los lectores que deseen mayores detalles acerca de estos temas deberían consultar AERA, APA, NCME (1999), Cole y Moss (1989), Cronbach (1988) y Jensen (1980, cap. 15).

EFECTOS COLATERALES INVOLUNTARIOS DE LAS PRUEBAS

El resultado deseado de la utilización de una prueba psicológica no necesariamente es la única consecuencia. También son posibles diversos efectos colaterales y, de hecho, son probables. El examinador debe determinar si los beneficios de aplicar una prueba superan los costos de los efectos colaterales potenciales. Lo que es más, al anticipar los efectos colaterales involuntarios, el examinador podría desviarlos o disminuirlos.

Cole y Moss (1989) citan el ejemplo del empleo de pruebas psicológicas para determinar la elegibilidad para educación especial. Aunque el resultado deseado es ayudar a los estudiantes a aprender, el proceso de identificar a los estudiantes como elegibles para educación especial puede producir numerosos efectos colaterales negativos:

- Los niños identificados pueden sentirse anormales o ineptos.
- Otros niños les pueden poner apodos.
- Los maestros pueden considerar que estos niños no merecen atención.
- El proceso puede producir salones segregados por raza o clase social.

La consideración de los efectos colaterales debería influir la decisión de un examinador de utilizar una prueba particular para un propósito específico. Éste podría elegir de manera apropiada no utilizar una prueba para un propósito meritorio, si los probables costos de los efectos colaterales superaran a los beneficios esperados.

Considérese la práctica común en años anteriores de utilizar el Inventario Multifásico de Personalidad de Minnesota (MMPI) para ayudar a seleccionar candidatos para puestos de imposición del orden público como oficiales de policía o ayudantes de comisario. Aunque el MMPI se diseñó originalmente como auxiliar en el diagnóstico psiquiátrico, la investigación subsecuente indicó que también era útil en la identificación de personas inadecuadas para una carrera en la policía (Hargrave, 1985; Hiatt y Hargrave, 1988). En particular, los agentes del orden público que produjeran perfiles en el MMPI con elevaciones leves (p. ej., puntuación T de 65 a 69) en las Escalas F (frecuencia), Masculinidad-Femineidad, Paranoia e Hipomanía tendían a estar implicados en graves acciones disciplinarias; los agentes del orden público que producían perfiles más "defensivos" en el MMPI, con elevaciones menores en las escalas clínicas, tendían a no participar en dichas acciones. Así, la prueba poseía una validez modesta para el digno propósito de seleccionar a los candidatos a la policía. Pero ninguna prueba, ni siquiera el muy respetado MMPI, es perfectamente válida. Se harán a un lado a algunos buenos solicitantes debido a que sus resultados en el MMPI son marginales.

Quizá su puntuación T en la Escala de Paranoia es de 66 o en la Escala de Hipomanía es de 68. En el MMPI, una puntuación T de 70 se considera con frecuencia como el límite superior del rango "normal".

Un efecto colateral involuntario de utilizar el MMPI para la valoración de los solicitantes a puestos de orden público es que los candidatos al trabajo que no tengan éxito con alguna institución pueden quedar señalados con una etiqueta patológica como psicópatas, esquizofrénicos o paranoides. Esta etiqueta puede surgir a pesar de los mejores esfuerzos del psicólogo asesor, quien es posible que nunca haya utilizado algún término peyorativo en el informe de evaluación del candidato. Por lo común, la etiqueta se aplica cuando los administradores en el departamento que hizo la referencia examinan el perfil del MMPI y ven que el candidato obtuvo su puntuación más alta en una escala con el horrendo título de Desviación psicopática, Esquizofrenia, Hipocondria o Paranoia. Por desgracia, la comunidad de seguridad pública puede ser una fraternidad muy cerrada. Es común que los jefes de policía y comisarios intercambien informes verbales acerca de sus solicitantes de empleo, de modo que una etiqueta peyorativa puede seguir al candidato de un lugar a otro, excluyendo de manera permanente al solicitante del ingreso a la profesión de seguridad pública. Las repercusiones no sólo son injustas para el candidato sino que también hacen surgir el espectro de demandas legales contra la institución y el psicólogo asesor. Si se consideran todos estos elementos, el psicólogo asesor puede encontrar preferible el empleo de una prueba técnicamente menos válida para el mismo propósito, en particular si el instrumento alternativo no produce estos efectos colaterales involuntarios.

La sensibilidad renovada hacia las cuestiones externas a la validez ha producido que varios teóricos en el campo de las pruebas amplíen su definición de validez de una prueba. En la siguiente sección se revisan estos recientes avances, advirtiendo al lector que aún queda por surgir un consenso final acerca de la naturaleza de la validez de una prueba.

AMPLIACIÓN DEL ESPECTRO DE VALIDEZ DE UNA PRUEBA

Para este momento, el lector se habrá familiarizado con la perspectiva estrecha y tradicionalista en el empleo de pruebas que afirma que ésta es válida si mide "lo que pretende medir". El mensaje implícito de esta perspectiva es que la validez técnica es la base más esencial para recomendar la utilización de una prueba. Después de todo, las pruebas válidas proporcionan información precisa acerca de las personas evaluadas —¿y qué podría haber de malo en ello?

De manera reciente, varios teóricos en psicometría han introducido una definición más amplia, funcionalista, de validez que asevera que una prueba es válida si satisface el propósito para el cual se utiliza (Cronbach, 1988; Messick, 1995). Por ejemplo, podría utilizarse una prueba de rendimiento en lectura para identificar a estudiantes a fin de asignarlos a la sección de regularización. Según la perspectiva funcionalista, la prueba sería válida —y, por tanto, su empleo sería apropiado— si los estudiantes seleccionados para regularización recibieran de hecho algún beneficio académico de esta aplicación de la prueba.

La perspectiva funcionalista reconoce de manera explícita que quien valida la prueba tiene la obligación de determinar si una práctica tiene consecuencias constructivas para los individuos e instituciones y, en especial, para proteger de los resultados adversos (Messick, 1980). Así, la validez de una prueba es un juicio valorativo general acerca de la idoneidad y pertinencia de las inferencias y **acciones** que fluyan de las puntuaciones de ésta.

Messick (1980, 1995) afirma que el concepto nuevo y más amplio de validez descansa sobre cuatro bases. Éstas son: 1) evidencia tradicional de validez de constructo; por ejemplo, validez convergente y de discriminación adecuadas, 2) un análisis de las implicaciones de valor de la interpretación de la prueba, 3) evidencia de la utilidad de las interpretaciones de una prueba en aplicaciones particulares y 4) una estimación de las consecuencias sociales potenciales y reales, incluyendo efectos colaterales,

del empleo de ésta. Una prueba válida es aquella que responde de manera apropiada a las cuatro facetas de validez.

La concepción más amplia de validez de una prueba es a todas luces polémica y algunos teóricos prefieren la perspectiva tradicional de que las consecuencias y valores son importantes, pero independientes de las cuestiones técnicas de su validez. Todos pueden estar de acuerdo con una cuestión: la medición psicológica no es un empeño neutral, es una ciencia aplicada que ocurre dentro de un contexto social y político.

RESUMEN

1. La validez de una prueba es el grado en el cual mide aquello que afirma medir. Una prueba es válida en el grado en que las inferencias que se realizan a partir de ella son apropiadas, significativas y útiles. La confiabilidad es un precursor necesario pero no suficiente de la validez.
2. Por tradición, las diferentes maneras de acumular evidencia de validez se han agrupado en tres categorías: validez de contenido, validez relacionada con el criterio y validez de constructo. Sin embargo, la validez es un concepto unitario y cualquier estudio empírico puede tener influencia sobre la validez de una prueba.
3. La validez de contenido se determina por el grado al que las preguntas, tareas o reactivos en una prueba constituyen una muestra representativa del universo de conductas que la prueba intenta examinar. Dicha validez es fácil de garantizar en el caso de rasgos bien definidos, como la capacidad ortográfica, pero es más difícil de especificar cuando se trata de rasgos explícitos, como la ansiedad.
4. Una prueba tiene validez de Facie si parece válida para los usuarios de la prueba, examinadores y, en especial, para las personas examinadas. La validez de Facie es importante para la aceptabilidad social de una prueba, pero es irrelevante para propósitos psicométricos.
5. La validez relacionada con el criterio se demuestra cuando una prueba es eficaz para predecir el desempeño en una medida apropiada de resultado. Dicha validez engloba a la concurrente, en la que las medidas de criterio se obtienen aproximadamente al mismo tiempo que las puntuaciones de la prueba que se utiliza como medio de predicción, y a la validez predictiva, en la que las medidas de criterio se obtienen en el futuro.
6. Cuando las pruebas se utilizan con propósitos de predicción, es necesario desarrollar una ecuación de regresión. Ésta describe la línea recta que se ajusta mejor (aquella que reduce al mínimo la suma de desviaciones cuadradas con respecto a la línea) para estimar el criterio a partir de la prueba. Por ejemplo, la ecuación $Y = .07X + .2$ podría utilizarse para predecir las calificaciones laborales a partir de una prueba de empleo.
7. La correlación entre prueba y criterio (r_{xy}) se conoce como coeficiente de validez. Mientras más alta sea la correlación, más precisa será la prueba en la estimación del criterio.
8. El error estándar del estimado (EE_{est}) es el margen de error a esperar en la puntuación de criterio predicha. Éste se obtiene de la siguiente fórmula:

$$EE_{est} = DE_y \sqrt{1 - r_{xy}^2}$$

donde r_{xy} es el coeficiente de validez.

9. Los defensores de la teoría de decisión destacan que una prueba debe auxiliar en la toma precisa de decisiones. La predicción precisa del éxito contra el fracaso en una medida de resultado es esencial. Las pruebas

deben evitar dos tipos de errores: los falsos positivos, en los que se pronostica que los individuos tendrán éxito, pero fracasan; y los falsos negativos, en los que se predice el fracaso de éstos, pero tienen éxito.

10. La teoría de decisión supone que los costos de las predicciones precisas e imprecisas pueden medirse en una escala común de utilidad, como la de ganancia/pérdida. Una suposición fundamental de la teoría de decisión es la maximización: En las decisiones institucionales de selección, la estrategia más apropiada de utilización de pruebas es aquella que aumenta al máximo la ganancia promedio y reduce al mínimo la pérdida promedio.
11. Un constructo es una cualidad o rasgo teórico, intangible, en el que los individuos difieren. La validez de constructo tiene que ver con las pruebas psicológicas que afirman medir atributos psicológicos complejos, multifacéticos y vinculados con la teoría como la capacidad de liderazgo, la hostilidad sobrecontrolada y la inteligencia.
12. Los estudios de validez de constructo caen por lo general dentro de una de estas categorías: análisis de la homogeneidad de los reactivos; evaluación de los cambios asociados con el desarrollo y de cambios grupales en la prueba; análisis de los efectos de la intervención; y correlación y análisis factoriales de las puntuaciones de una prueba en relación con otras fuentes de información. En cada caso, la pregunta crucial es si los resultados son consistentes con la teoría subyacente del constructo que se mide.
13. Las preocupaciones externas a la validez incluyen los efectos colaterales y consecuencias involuntarias de las pruebas. Por ejemplo, una evaluación válida para asignación a educación especial puede causar que los niños identificados se sientan anormales o ineptos. La consideración de los efectos colaterales debería influir sobre la decisión del examinador de utilizar una prueba particular para un propósito específico.
14. La perspectiva nueva, más amplia y funcionalista de la validez de una prueba afirma que ésta es válida si satisface el propósito para el cual se utiliza. Por ejemplo, la validez de una prueba de rendimiento en lectura podría vincularse con la regularización exitosa de los estudiantes con discapacidad para la lectura que se hayan identificado con la prueba.

TÉRMINOS Y CONCEPTOS CLAVE

Validez p. 118
Validez de contenido p. 119
Validez de Facie p. 122
Validez relacionada con el criterio p. 122
Validez concurrente p. 124
Validez predictiva p. 125
Ecuación de regresión p. 125
Coeficiente de validez p. 125
Error estándar del estimado p. 125
Teoría de decisión p. 126
Falsos positivos p. 127
Falsos negativos p. 127
Tasa base p. 129
Constructo p. 131
Validez de constructo p. 131
Escala homogénea p. 132
Validez convergente p. 135
Validez de discriminación p. 135
Matriz multirrasgo-multimétodo p. 135
Análisis factorial p. 136
Carga factorial p. 136
Preocupaciones externas a la validez p. 137

TEMA 4B
Construcción de pruebas

Definición de la prueba
Selección de un método de escalamiento
Métodos representativos de escalamiento
Construcción de reactivos
Examen de reactivos
Revisión de la prueba
Publicación de la prueba
Resumen
Términos y conceptos clave

La creación de una nueva prueba implica tanto una ciencia como un arte. Un autor de pruebas debe elegir estrategias y materiales y después tomar decisiones continuas de investigación que afectarán la calidad de su instrumento en germinación. El propósito de esta sección consiste en analizar el proceso mediante el cual los psicómetras creen pruebas válidas. Aunque se examinarán muchas cuestiones independientes, a todas las vincula un tema en común: las pruebas válidas simplemente no aparecen en escena en plena madurez —surgen de manera lenta de un proceso evolutivo, en desarrollo, que se funda en la validez desde el mismo principio. En este tema se enfatizarán los aspectos básicos del desarrollo de pruebas; el lector que desee una presentación más avanzada deberá consultar a Kline (1986) y Nunnally (1978).

La construcción de pruebas consta de seis etapas entrelazadas:

Definición de la prueba.
Selección de un método de escalamiento.
Construcción de los reactivos.
Examen de los reactivos.
Revisión de la prueba.
Publicación de la prueba.

Como análisis preliminar, pueden resumirse estos pasos de la siguiente manera: la definición de la prueba consiste en la delimitación de su alcance y propósito, lós cuales deben conocerse antes de que el autor pueda proseguir con su construcción. La selección de un método de escalamiento es un proceso en el que se establecen las reglas mediante las cuales se asignan números a los resultados de la prueba. La construcción de los reactivos es tanto un arte como una ciencia y es aquí donde puede requerirse la creatividad del autor. Una vez que se tiene disponible una versión preliminar de la prueba, los autores en general la aplican a una muestra de tamaño modesto de individuos, a fin de reunir los datos iniciales sobre las características de los reactivos. El examen de éstos implica una variedad de procedimientos estadísticos conocidos como **análisis de reactivos**, cuyo propósito consiste en determinar cuáles reactivos deben conservarse, revisarse y eliminarse. Con base en el análisis de reactivos y en otras fuentes de información, se revisa la prueba. Si las revisiones son sustanciales, es posible que se requieran nuevos reactivos y prepruebas adicionales con nuevos individuos. Así, la construcción de pruebas implica un circuito de realimentación donde se podrían producir 2, 3 y 4 bosquejos de un instrumento (figura 4–5). La publicación de la prueba es el paso final. Además de difundir los materiales de una prueba, el autor debe producir un manual fácil de comprender para el usuario. A continuación se examinarán con más detalle cada uno de estos pasos.

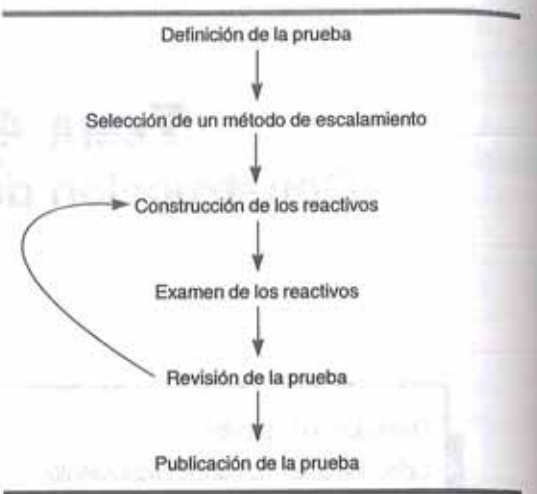


Figura 4–5. Procesos de construcción de la prueba.

DEFINICIÓN DE LA PRUEBA

A fin de construir una nueva prueba, el autor debe tener una idea clara de aquello que ha de medir y de cómo difiere ésta de los instrumentos existentes. Dado que la existencia de las pruebas psicológicas está a punto de entrar en su segundo siglo y ya que se han publicado miles de pruebas para este momento, es claro que la responsabilidad de probar que un instrumento es diferente y mejor que los existentes depende del autor.

Considérese la sobrecogedora tarea que enfrenta un autor de pruebas que propone otro instrumento de inteligencia general. Con docenas de dichos instrumentos en existencia, ¿cómo es posible que uno nuevo haga una contribución útil al campo? La respuesta es que la investigación contemporánea aumenta de manera continua la comprensión acerca de la inteligencia e impulsa a buscar nuevas maneras y más útiles de medir este constructo multifacético.

Kaufman y Kaufman (1983) proporcionan un buen modelo del proceso de definición de la prueba. Al proponer la Batería de Evaluación de Kaufman para Niños (K-ABC), una nueva prueba de la inteligencia general en niños, los autores enumeraron seis objetivos primarios que definen el propósito de la prueba y la distinguen de las medidas existentes:

1. Mide la inteligencia desde una fuerte base teórica y de investigación.
2. Separa el conocimiento objetivo adquirido de la capacidad para resolver problemas desconocidos.
3. Produce puntuaciones que se traducen en intervención educativa.
4. Incluye tareas novedosas.

5. Es fácil de aplicar y objetiva de calificar.
6. Es sensible a las diversas necesidades de los preescolares, grupos minoritarios y niños excepcionales (Kaufman y Kaufman, 1983).

Como el lector descubrirá en un tema posterior, la K-ABC representa un viraje interesante con respecto a las pruebas tradicionales de inteligencia. Por ahora, la cuestión importante es que los autores de este instrumento reciente explicaron su propósito de manera explícita y plantearon un nuevo enfoque para medir la inteligencia, mucho antes de que comenzaran a construir los reactivos de la prueba.

SELECCIÓN DE UN MÉTODO DE ESCALAMIENTO

El propósito inmediato del examen psicológico es asignar números a las respuestas en una prueba de modo que pueda juzgarse que la persona examinada tiene una mayor o menor cantidad de la característica medida. Las reglas mediante las cuales se asignan números a las respuestas definen el método de escalamiento. Los autores de pruebas seleccionan un método de este tipo que se adecue en forma óptima a la manera en que han conceptuado el rasgo o rasgos medidos por su prueba. Ningún método de escalamiento es uniformemente mejor que otros. En el caso de algunos rasgos, la clasificación ordinal de parte de jueces expertos podría ser el mejor enfoque de medición; en el caso de otros, la medición escalar compleja de los datos de autoinforme puede producir las medidas más válidas.

Existen tantos métodos precisos de escalamiento disponibles para los psicómetras que nos sentiremos satisfechos con proporcionar aquí sólo una muestra representativa. Los lectores que deseen tener una reseña más completa y detallada deben consultar a Gulliksen (1950), Nunnally (1978) o Kline (1986). No obstante, antes de revisar los métodos de escalamiento seleccionados, es necesario que se presente un concepto relacionado, los niveles de medición, de modo que el lector pueda apreciar mejor las diferencias entre los métodos de escalamiento.

NIVELES DE MEDICIÓN

Según Stevens (1946) todos los números extraídos de los instrumentos de medición de cualquier tipo pueden colocarse dentro de una de cuatro categorías jerárquicas: nominal, ordinal, de intervalo o de razón. Cada categoría define un nivel de medición; el orden que se lista es de menos a más informativo.

En las **escalas nominales**, los números sirven sólo como nombres de categorías. Por ejemplo, cuando se obtienen datos para un estudio demográfico, un investigador podría codificar a los varones como "1" y a las mujeres como "2". Véase que los números son arbitrarios y no designan si hay "más" o "menos" de una cosa. En las escalas nominales, los números son sólo una forma simplificada de denominación.

Las **escalas ordinales** constituyen una forma de ordenación o clasificación. Si se le pidiera a profesores universitarios que clasificaran cuatro automóviles en cuanto a cuál preferirían poseer, el orden preferido podría ser "1" *Cadillac*, "2" *Chevrolet*, "3" *Volkswagen*, "4" *Hyundai*. Obsérvese en este caso que los números no son intercambiables. Una clasificación de "1" es "más" que una clasificación de "2" y así sucesivamente. El "más" se refiere al orden de preferencia. Sin embargo, las escalas ordinales no proporcionan información sobre la fortaleza relativa de las clasificaciones. En este ejemplo hipotético no se puede saber si los profesores universitarios tienen una fuerte preferencia por los *Cadillac* por encima de los *Chevrolet* o sólo hay una diferencia marginal en su preferencia.

Una **escala de intervalo** proporciona información acerca de clasificación, pero también provee una medida para estimar las diferencias entre clasificaciones. Para construir una escala de intervalo se podría pedir a los profesores universitarios que calificaran en una escala de 1 a 100 qué tanto les gustaría poseer los cuatro automóviles antes citados. Supóngase que las

calificaciones promedio van como sigue: *Cadillac*, 90; *Chevrolet*, 70; *Volkswagen*, 60; *Hyundai*, 50. A partir de esta información se puede inferir que la preferencia por un *Cadillac* es mucho más fuerte que por un *Chevrolet* que, a su vez, tiene una preferencia levemente más fuerte que el *Volkswagen*. Más importante aún es que también puede suponerse que los intervalos entre los puntos en esta escala son aproximadamente iguales: La diferencia entre la preferencia de los profesores por un *Chevrolet* y un *Volkswagen* (10 puntos) es más o menos la misma que existe entre un *Volkswagen* y un *Hyundai* (también 10 puntos). En pocas palabras, las escalas de intervalo se basan en la suposición de unidades o intervalos de igual tamaño para la escala subyacente.

Una **escala de razón** tiene todas las características de una escala de intervalo, pero también posee un punto cero conceptualmente significativo, en el que existe ausencia total de la característica medida. Las propiedades esenciales de los cuatro niveles de medición se resumen en la figura 4–6.

Las escalas de razón son poco comunes en la medición psicológica. Considérese si existe algún sentido significativo en el que pueda pensarse que una persona tiene cero inteligencia. No, en realidad. Lo mismo ocurre con la mayor parte de los constructos en psicología: los puntos cero significativos simplemente no existen. Sin embargo, unas cuantas medidas físicas que emplean los psicólogos califican para escalas de razón. Por ejemplo, la estatura y el peso alcanzan esa calificación y quizá también algunas medidas fisiológicas como la respuesta electrodérmica. Pero, en general, lo más que puede esperar un psicólogo es una media al nivel de intervalo.

	Características			
Nivel	Permite la categorización	Permite la clasificación	Utiliza intervalos iguales	Posee un punto cero real
Nominal	X			
Ordinal	X	X		
De intervalo	X	X	X	
De razón	X	X	X	X

Figura 4–6. Características esenciales de los cuatro niveles de medición.

Los niveles de medición son relevantes para la construcción de pruebas debido a que los procedimientos estadísticos paramétricos más poderosos y útiles (p. ej., *r* de Pearson, análisis de varianza, regresión múltiple) deberían utilizarse sólo para puntuaciones obtenidas de medidas que satisfacen los criterios de las escalas de intervalo o de razón. Para las escalas que sólo son nominales u ordinales deben emplearse procedimientos estadísticos no paramétricos menos poderosos (p. ej., ji cuadrada, correlación por rangos, pruebas de mediana). En la práctica, se supone que la mayor parte de los principales instrumentos de examen psicológico (en especial las pruebas de inteligencia y escalas de personalidad) emplean una medición que de manera aproximada está al nivel de intervalo aunque, en términos estrictos, es muy difícil demostrar igualdad absoluta de intervalos para dichos instrumentos (Bausell, 1986). Ahora que el lector se ha familiarizado con los niveles de medición, se presenta una muestra representativa de los métodos de escalamiento, señalando de antemano que los diferentes métodos producen distintos niveles de medición.

MÉTODOS REPRESENTATIVOS DE ESCALAMIENTO

CLASIFICACIONES DE EXPERTOS

Supóngase que se desea medir la profundidad del coma en los pacientes que han sufrido un traumatismo craneal reciente que los ha dejado inconscientes. Una escala de profundidad del coma sería muy importante para la predicción del curso de la mejoría, debido a que es bien conocido que un periodo extenso de inconsciencia ofrece un pronóstico pobre de recuperación. Además, el personal de rehabilitación tiene la necesidad práctica de saber si un paciente está en coma profundo o se encuentra en un estado

parcialmente comunicativo de conciencia con somnolencia.

Un enfoque de la medición con una escala de la profundidad del coma consistiría en depender de las **clasificaciones de expertos** en cuanto a aspectos conductuales. Por ejemplo, se le puede pedir a un panel de neurólogos que enumere las conductas asociadas con los diferentes niveles de conciencia. Después de que éstos han presentado una gran lista de conductas diagnósticas, los autores de la prueba —preferiblemente expertos en traumatismos craneales— clasificarían las conductas indicadoras en un continuo de conciencia que vaya desde el coma profundo hasta la orientación básica. Teasdale y Jennett (1974) utilizaron justo este enfoque y produjeron la *Glasgow Coma Scale* (Escala de Coma de Glasgow). Instrumentos similares a éste se emplean de manera amplia en hospitales para la evaluación de lesiones por traumatismo craneal (figura 4–7).

La *Glasgow Coma Scale* se califica al observar al paciente y asignar el máximo nivel de funcionamiento en cada una de las tres subescalas. En cada una se supone que el paciente presenta todos los niveles de conducta por debajo del nivel calificado. Así, desde un punto de vista psicométrico, esta escala consiste de tres subescalas (ojos, respuesta verbal y respuesta motora) cada una de las cuales produce una clasificación ordinal de conducta.

Además de las clasificaciones, es posible calcular una sola puntuación total que se encuentra un poco más allá de una escala ordinal, aunque probablemente sea menos que una verdadera medición al nivel de intervalo. Si se asignan números a las clasificaciones (p. ej., para ojos abiertos una codificación de "no" = 1; "ante el dolor" = 2 y así sucesivamente), entonces pueden sumarse los números para el nivel calificado en cada subescala, lo cual produce una puntuación máxima posible de 14 puntos. La puntuación total en la *Glasgow Coma Scale* predice la recuperación posterior con un grado muy alto de precisión (Jennett, Teasdale y Knill-Jones, 1975). Por tanto, puede verse que las

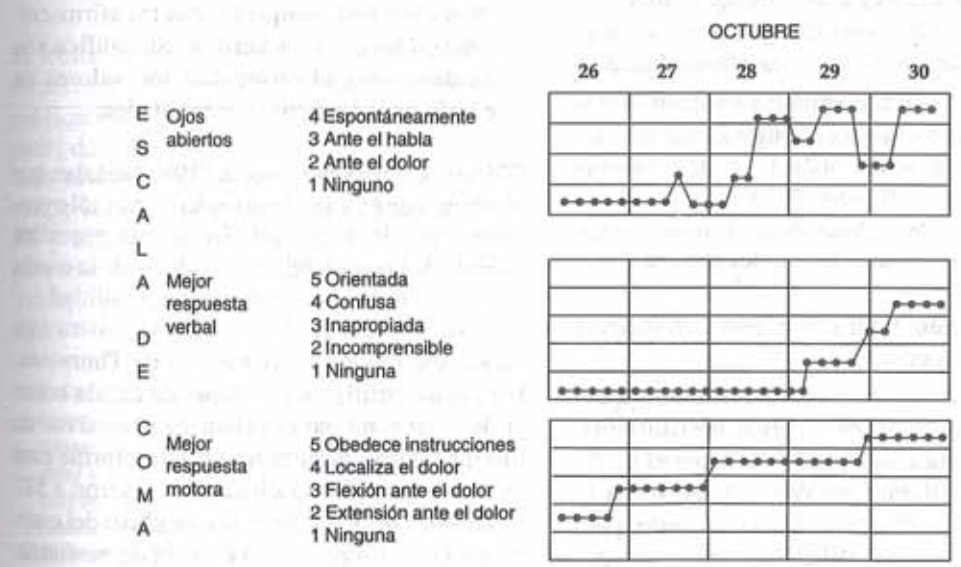


Figura 4–7. Ejemplo de la utilización de la Escala de Coma de Glasgow para el registro de la profundidad del coma.

Fuente: Reproducido con autorización de Jennett, B., Teasdale, G. M., and Knill-Jones, R. P. (1975). Predicting outcome after head injury. *Journal of the Royal College of Physicians of London, 9*, 231-237.

pruebas psicológicas bastante elementales obtenidas de los métodos de escalamiento más sencillos pueden, a pesar de ello, proporcionar información válida y útil.

MÉTODO DE INTERVALOS DE APARICIÓN BALANCEADA

A principios de este siglo, L. L. Thurstone (1929) propuso un método para construir escalas al nivel de intervalo a partir de afirmaciones de actitud. Su **método de intervalos de aparición balanceada** todavía se utiliza en la actualidad, lo cual hace que Thurstone sea uno de los gigantes de la teoría psicométrica. La metodología en sí para construir intervalos de aparición balanceada es un tanto estadísticamente compleja y abrumadora, pero la lógica subyacente es fácil de explicar (Ghiselli, Campbell y Zedeck, 1981). Para ilustrar este método, se resumen los pasos implicados en la construcción de una escala de actitudes hacia la membresía a una iglesia.

1. Se obtienen tantas afirmaciones posibles de falso-verdadero que reflejen una variedad de actitudes positivas y negativas hacia la iglesia. Dos ejemplos extremos podrían ser:

 "Siento que los servicios religiosos me dan inspiración y me ayudan a dar lo mejor de mí mismo durante la siguiente semana".

 "Pienso que las iglesias buscan imponer gran cantidad de dogmas anticuados y supersticiones medievales".

 Por supuesto, también se reunirían muchos reactivos moderados.
2. A continuación, se le pide a un número aproximado de 10 jueces expertos que califiquen estas afirmaciones para determinar el grado de favorabilidad/desfavorabilidad hacia la actitud. Los jueces deben ser calificados para esa tarea; podría utilizarse a ministros para una escala de actitud hacia la membresía a una iglesia. En general, se pide a cada juez que clasifique cada afirmación en categorías de 1 a 11 que vayan desde "extremadamente favorable" hasta "extremadamente desfavorable". Se le dice a los jueces que hagan a un lado sus propios prejuicios y que consideren las 11 categorías como equidistantes.
3. Después de que los jueces han terminado con el proceso de evaluación, se determina la calificación media de favorabilidad (de 1 a 11) y la desviación estándar para cada reactivo. Por ejemplo, 10 jueces pueden haber dado una calificación promedio de favorabilidad de 9.2 al primer reactivo antes enunciado; pero es probable que las calificaciones difieran de un juez a otro, como se refleja en una desviación estándar de 1.1 para este reactivo.
4. Debido a que la desviación estándar en la calificación de favorabilidad de un reactivo refleja ambigüedad, los reactivos con desviaciones estándar grandes se descartan. Por lo general, se eligen aproximadamente 20 a 30 reactivos, de modo que las afirmaciones cubran el rango de la dimensión (favorable a desfavorable). Se supone que las diferencias entre reactivos en la escala final satisfacen las propiedades de una escala de intervalo.
5. Se pide a las personas que contesten la escala de actitud que marquen todas las afirmaciones con las que concuerdan. Su calificación se determina al promediar los valores de escala de los reactivos respaldados.

Ghiselli, Campbell y Zedeck (1981) señalan que el método de escalamiento anterior tan sólo produce la escala de actitud. Todavía se necesitan análisis de la confiabilidad y validez de la escala a fin de determinar su idoneidad y utilidad.

Un estudio de Russo (1994) ilustra una aplicación moderna del método de Thurstone. Esta autora utilizó un enfoque de escala como el de Thurstone para evaluar 216 reactivos de tres inventarios destacados de autoinforme para evaluar depresión. Los jueces incluyeron a 527 estudiantes de pregrado y 37 miembros del cuerpo docente clínico de una escuela de medicina. Los 216 reactivos se distribuyeron en forma aleatoria y se calificaron con respecto a gravedad de la depresión desde 1, que representaba ausencia de depresión, hasta 11, que significaba depresión extrema. Russo descubrió que los

tres inventarios de autoinforme carecían de reactivos y opciones de respuesta típicos de la depresión leve. La distribución de los 216 reactivos fue bimodal, con muchos de ellos agrupados cerca de la parte baja (ninguna depresión) y muchos otros cerca de la parte media (depresión moderada). Una conclusión característica para uno de los conjuntos de reactivos de una destacada escala de depresión fue la siguiente:

Calificación para la depresión	Puntuación original	Contenido del reactivo
1.0	1	Nunca me siento melancólico o triste.
3.4	2	En ocasiones me siento melancólico o triste.
4.1	3	Me siento melancólico o triste una buena parte del tiempo.
4.4	4	Me siento melancólico o triste la mayor parte del tiempo.

El lector verá que la calificación original de estos reactivos se desvía sustancialmente de las calificaciones de depresión que proporcionó el panel de estudiantes y docentes clínicos. También es evidente que los valores reales de la escala son discontinuos, ya que van de 1.0 a 3.4 y más alto. Se observó un patrón similar para muchos de los reactivos en los tres inventarios, lo cual condujo a Russo (1994) a concluir:

> Estos resultados sugieren que si se utiliza la puntuación original para las tres escalas que se examinan aquí, entonces las distinciones entre bienestar y ausencia de depresión, al igual que entre depresión moderada y grave, serán difíciles de realizar. Tal imprecisión hará difícil la evaluación de la eficacia de los tratamientos para la depresión, debido a que una carencia como ésta debe estar en función de un error adicional de medición debido a las medidas ordinales. Tal error también crearía un desastre en estudios longitudinales, en especial en aquéllos en los que participa la memoria.

En este ejemplo puede verse que el enfoque de Thurstone para la asignación de una escala para los reactivos tiene poderosas aplicaciones en el desarrollo de pruebas. Con base en estos descubrimientos, los investigadores están ahora en posición de desarrollar escalas mejoradas de autoinforme que evalúen el rango completo de síntomas de la depresión.

MÉTODO DE ESCALAMIENTO ABSOLUTO

Thurstone (1925) también desarrolló el **método de escalamiento absoluto**, un procedimiento para obtener una medida de la dificultad absoluta de los reactivos con base en los resultados de diferentes grupos de personas examinadas de diferentes edades. La metodología para determinar la dificultad individual de los reactivos en un escalamiento absoluto es bastante compleja, aunque la fundamentación subyacente no es tan difícil de comprender. En esencia, se aplica un conjunto de reactivos comunes de prueba a dos o más grupos de diferentes edades. La dificultad relativa de éstos sirve como fundamento para realizar una serie de comparaciones entrelazadas para todos los reactivos y todos los grupos por edad. Un grupo por edad sirve como grupo base. La dificultad de los reactivos se mide en unidades comunes como unidades de capacidad en desviación estándar para el grupo base. El método de escalamiento absoluto se utiliza ampliamente en pruebas de rendimiento y aptitud (STEP, 1980; Donlon, 1984).

Thurstone (1925) ilustró el método de escalamiento absoluto con datos de prueba de 3 000 escolares en las 65 preguntas de la prueba Binet original. Con la media de inteligencia de los niños de 3 años, 6 meses de edad en la prueba Binet como punto cero y la desviación estándar de su inteligencia como unidad de medición, construyó una escala que iba de –2 a +10 y después localizó cada una de las 65 preguntas en esa escala. Thurstone (1925) encontró que la escala "hace surgir de manera bastante sorprendente el hecho de que las preguntas se encuentran agrupadas de manera indebida en

ciertos rangos [de dificultad] y son más bien escasas en otros". Un autor moderno de pruebas utilizaría este tipo de análisis como base para eliminar reactivos redundantes (en el sentido de que miden el mismo nivel de dificultad) y añadiría otros que analicen los rangos superiores (e inferiores) de dificultad.

ESCALAS LIKERT

Likert (1932) propuso un método sencillo y directo para asignar una escala a las actitudes, el cual tiene amplia utilización hoy día. Una **escala Likert** le presenta a la persona evaluada cinco respuestas ordenadas en un continuo de acuerdo/desacuerdo o aprobación/desaprobación. Por ejemplo, un reactivo en una escala para evaluar las actitudes hacia la membresía a una iglesia podría ser:

Los servicios religiosos me dan inspiración y me ayudan a dar lo mejor de mí mismo durante la siguiente semana.

Usted está:

Fuertemente de acuerdo	De acuerdo	Indeciso	En desacuerdo	Fuertemente en desacuerdo

Dependiendo del fraseo de un reactivo en particular, una respuesta extrema de "fuertemente de acuerdo" o "fuertemente en desacuerdo" indicaría la respuesta más favorable en la actitud subyacente medida con este cuestionario. Likert (1932) asignó una puntuación de 5 a esta respuesta extrema, 1 a la respuestas totalmente contraria y 2, 3 y 4 a las respuestas intermedias. La puntuación total de la escala se obtiene al sumar las puntuaciones de los reactivos individuales. Por tal razón, la escala Likert también se conoce como **escala sumatoria.**

ESCALAS GUTTMAN

En una **escala Guttman,** las personas que corroboran una afirmación también concuerdan con afirmaciones más leves que tienen que ver con el mismo continuo subyacente (Guttman, 1944, 1947). Así, cuando el examinador conoce la confirmación más extrema de la persona evaluada en el continuo, le es posible reconstruir también la respuesta intermedia. Las escalas Guttman se producen al seleccionar reactivos que caigan en una secuencia ordenada de confirmaciones de la persona examinada. Es poco común que se logre una escala Guttman perfecta debido a los errores de medición pero, de todas maneras, es una meta adecuada para ciertos tipos de pruebas.

Aunque el enfoque Guttman se diseñó originalmente para determinar si un conjunto de afirmaciones de actitud es unidimensional, la técnica se ha utilizado para muchos tipos diferentes de pruebas. Por ejemplo, Beck utilizó un método de escalamiento del tipo Guttman para producir los reactivos individuales del *Beck Depression Inventory* (Inventario Beck de Depresión; BDI, Beck, Steer y Garbin, 1988; Beck y colaboradores, 1961). Los reactivos de este inventario se parecen a los siguientes:

() En ocasiones me siento triste o afligido
() Con frecuencia me siento triste o afligido
() Me siento triste o afligido la mayor parte del tiempo
() Siempre me siento triste y no puedo tolerarlo.

Se pide al cliente que "marque en cada grupo las afirmaciones que, a su parecer, lo representen de manera más certera". Un cliente que marque una alternativa extrema (p. ej., "Siempre me siento triste y no puedo tolerarlo") casi con toda seguridad coincidirá también con las afirmaciones más leves.

MÉTODO DE CONCORDANCIA EMPÍRICA

Es posible que el lector haya observado que la mayor parte de los métodos de escalamiento que se analizan en la sección anterior dependen del juicio autorizado de expertos para la selección

y ordenación de los reactivos. También es posible construir escalas de medición que se basen por completo en consideraciones empíricas ajenas a la teoría o el juicio experto. En el **método de concordancia empírica**, los reactivos de una prueba se seleccionan para una escala con base únicamente en qué tan bien contrastan con un grupo criterio con respecto a una muestra normativa. Por ejemplo, se podría obtener una escala de Depresión a partir de una reserva de preguntas de falso-verdadero de un inventario de personalidad, como se ve a continuación:

1. Se reúne a un grupo cuidadosamente seleccionado y homogéneo de personas que experimenten depresión mayor, a fin de que respondan a la reserva de preguntas de falso-verdadero.
2. En cada reactivo se compara la frecuencia de confirmación del grupo de depresión con la de la muestra normativa.
3. Los reactivos que muestran una gran diferencia en frecuencia de confirmación entre las muestras con depresión y normativa se seleccionan para la escala de Depresión, se codifican en la dirección favorecida por los individuos con depresión (verdadero o falso, según sea apropiado).
4. Entonces, la puntuación en bruto de la escala de Depresión es sencillamente el número de reactivos respondidos en la dirección codificada.

El método de concordancia empírica puede producir algunas sorpresas interesantes. Un descubrimiento común es que algunos reactivos seleccionados para una escala pueden no mostrar relación obvia con el constructo a medir. Por ejemplo, un reactivo como "Bebo mucho agua" (codificado como verdadero) podría terminar dentro de una escala de Depresión. La fundamentación momentánea para incluir este reactivo es sencillamente que funciona. Por supuesto, el reto que se le presenta a los investigadores consiste en determinar por qué funciona el reactivo. Sin embargo, desde el punto de vista práctico de la construcción de una escala empírica, las consideraciones teóricas tienen importancia secundaria. En el tema 14A, Inventarios de autoinforme, se analiza con mayor amplitud el método de concordancia empírica.

CONSTRUCCIÓN DE ESCALAMIENTO RACIONAL (CONSISTENCIA INTERNA)

El enfoque racional de la construcción de una escala es un método popular para el desarrollo de los inventarios de autoinforme. El nombre **racional** es en cierto modo equivocado, ya que algunos métodos estadísticos son esenciales para este enfoque. También, el nombre implica que otros enfoques no son racionales o son irracionales, lo cual es falso. En el centro del **método de escalamiento racional** se encuentra que todos los reactivos de la escala se correlacionan en sentido positivo entre sí y también con la puntuación total de la misma. Un nombre alternativo y más apropiado para este enfoque es el de consistencia interna, el cual enfatiza lo que en realidad se hace. Gough y Bradley (1992) explican cómo fue que el enfoque racional obtuvo su descriptivo título:

> La idea de la racionalidad entra en escena en el momento en que el tema central o dimensión unificadora, alrededor del cual se agrupan los reactivos, es aquel que de antemano articuló conceptualmente el autor de la medida y a partir del cual se determina la calificación de cada reactivo de manera lógica y comprensible.

Se seguirá su presentación para ilustrar los aspectos del enfoque racional.

Supóngase que un autor de pruebas desea desarrollar una nueva escala de autoinforme para el potencial de liderazgo. Con base en una revisión de la literatura pertinente, el investigador podría concluir que el potencial de liderazgo se caracteriza por confianza en uno mismo, flexibilidad bajo presión, inteligencia elevada, capacidad de persuasión, asertividad y la capacidad para percibir lo que otros piensan y sienten. Estos conceptos sugieren que los

siguientes reactivos de falso-verdadero podrían ser útiles en la evaluación del potencial de liderazgo (Gough y Bradley, 1992):

- En general tengo confianza y seguridad en mí mismo. (V)
- Cuando otras personas están en desacuerdo conmigo, en general me quedo en silencio o cedo de algún otro modo. (F)
- Creo que me encuentro notablemente por encima del promedio en cuanto a capacidad intelectual. (V)
- Con frecuencia siento que tengo una comprensión deficiente de la manera en que otras personas reaccionarán ante las cosas. (F).
- Es probable que mis amigos me describieran como una persona fuerte y enérgica. (V)

La **V** y la **F** después de cada afirmación indican la dirección codificada de manera racional para el potencial de liderazgo.

Por supuesto, también se propondrían reactivos adicionales con intenciones similares. El autor de la prueba podría comenzar con 100 reactivos que parecen —con una base racional— evaluar el potencial de liderazgo. Estos reactivos preliminares se aplicarían a una muestra amplia de individuos similares a la población blanco a la que se dirige la escala. Por ejemplo, si ésta se diseña para identificar a estudiantes universitarios con potencial de liderazgo, entonces debería aplicarse a una muestra representativa de varios cientos de estudiantes universitarios. Para el desarrollo de la escala son deseables muestras muy grandes. En este caso hipotético, supóngase que se obtienen resultados con 500 estudiantes universitarios.

El siguiente paso en la construcción de un escalamiento racional consiste en correlacionar las puntuaciones de cada uno de los reactivos preliminares con la puntuación total de la prueba para los 500 participantes en la muestra del piloteo. Debido a que las puntuaciones en los reactivos son dicotómicas (se asigna de manera arbitraria un 1 para la respuesta que corresponde a la clave de calificación y 0 a la alternativa), se requiere de un coeficiente r_{bs} de correlación biserial. Una vez que se obtienen las correlaciones, el investigador busca en el listado las correlaciones débiles y las inversas (correlaciones negativas). Estos reactivos se descartan debido a que no contribuyen a la medición del potencial de liderazgo. Podrían eliminarse hasta la mitad de los reactivos originales. Si de inicio se descarta una gran proporción de éstos, el investigador podría volver a calcular las correlaciones entre los reactivos y el total, con base en una reserva reducida de aquéllos para verificar la homogeneidad de los restantes. Los reactivos que sobreviven a este procedimiento iterativo constituyen la escala de potencial de liderazgo. El lector debería mantener en mente que el enfoque racional para la construcción de escalas tan sólo produce una homogénea al medir un constructo especificado. Se necesitarían estudios adicionales con nuevas muestras de individuos para determinar la confiabilidad y validez de la nueva escala.

CONSTRUCCIÓN DE REACTIVOS

Construir los reactivos de una prueba es un procedimiento doloroso y laborioso que impone demandas a la creatividad de los autores de pruebas. Quien los redacta se enfrenta con una profusión de preguntas iniciales:

- ¿El contenido de los reactivos debe ser homogéneo o variado?
- ¿Qué rango de dificultad deberían cubrir los reactivos?
- ¿Cuántos reactivos iniciales deberían construirse?
- ¿Qué procesos cognitivos y dominios del reactivo deberían cubrirse?
- ¿Qué tipos de reactivos de prueba deberían utilizarse?

Se atenderá de manera breve a las tres primeras preguntas antes de volver a un análisis más detallado de los dos últimos temas, que por lo común se engloban bajo los encabezados de tabla de especificaciones y formatos de reactivo.

PREGUNTAS INICIALES EN CONSTRUCCIÓN DE PRUEBAS

La primera pregunta tiene que ver con la homogeneidad contra la heterogeneidad en el contenido de los reactivos de una prueba. En gran medida, aquello que dicta si el contenido de los reactivos ha de ser homogéneo o variado es la manera en que el autor de la prueba ha definido al nuevo instrumento. Considérese una prueba culturalmente reducida de la inteligencia general. Dicho instrumento podría incorporar reactivos variados, en tanto las preguntas no supongan una escolaridad específica. El autor de la prueba podría tener la intención de incorporar problemas novedosos que sean igualmente desconocidos para todas las personas evaluadas. Por otro lado, con una prueba de pensamiento espacial que se base en la teoría, se requerirían subescalas con reactivos que tengan contenido homogéneo.

El rango de dificultad de los reactivos debe ser el suficiente como para dar lugar a una diferenciación significativa de las personas evaluadas en ambos extremos. Entonces, las pruebas más útiles son aquellas que incluyen una serie graduada de reactivos muy fáciles, que aprueban casi todos, al igual que un grupo de éstos cada vez más difíciles que casi nadie aprueba. Se observa un efecto de límite superior cuando números significativos de examinandos obtienen puntuaciones perfectas o casi perfectas. El problema con el efecto de límite superior es que no son posibles las distinciones entre personas con puntuaciones elevadas, aunque estos individuos podrían diferir de manera sustancial en cuanto al rasgo subyacente que mide la prueba. Se observa un efecto de límite inferior cuando cantidades significativas de examinandos obtienen puntuaciones que se encuentran cerca o en la parte más baja de la escala. Por ejemplo, el WAIS-R tiene un importante efecto de límite inferior en cuanto a que no puede discriminar entre los niveles moderado, grave y profundo de retraso mental —todas las personas con discapacidades significativas del desarrollo fallan en sus respuestas a casi todas las preguntas.

Los autores de pruebas esperan que algunos de los reactivos iniciales produzcan contribuciones poco eficaces para el objetivo global de medición de su instrumento. Por esta razón, es común la práctica de construir un primer borrador que contenga un exceso de reactivos, quizás el doble del número de preguntas que se deseen para el bosquejo final. Por ejemplo, el MMPI, que contiene 550 reactivos, originalmente consistía de más de 1 000 afirmaciones de falso-verdadero acerca de la personalidad (Hathaway y McKinley, 1940).

TABLA DE ESPECIFICACIONES

Es frecuente que los autores profesionales de pruebas de rendimiento y capacidad utilicen uno o más esquemas de redacción de reactivos para ayudar a garantizar que su instrumento detecte una mezcla deseada de procesos cognitivos y dominios de contenido. Por ejemplo, un esquema muy sencillo de redacción de reactivos podría establecer que una prueba de rendimiento sobre la Guerra Civil de los Estados Unidos debería consistir de 10 reactivos de opción múltiple y 10 preguntas de llenado de espacios en blanco, la mitad de ellos sobre cuestiones objetivas (p. ej., fechas, principales batallas) y la otra mitad sobre asuntos conceptuales (p. ej., diferentes perspectivas acerca de la esclavitud).

Antes de que comience el desarrollo de la prueba, por lo general los redactores de los reactivos reciben una **tabla de especificaciones**. Ésta enumera la información y las tareas cognitivas en las que se ha de evaluar a los examinandos. Quizás la tabla más común sea la matriz de contenido por proceso, que lista el número exacto de reactivos en las áreas pertinentes de contenido y detalla la combinación precisa de reactivos que deben ejemplificar diferentes procesos cognitivos (Millman y Greene, 1989).

Considérese una prueba de rendimiento en ciencias naturales adecuada para estudiantes de educación media. Dicha prueba debería cubrir muchas áreas diferentes de contenido y debería requerir una mezcla de procesos cognitivos

Cuadro 4–6. Ejemplo de una tabla de especificaciones de contenido por proceso para una prueba hipotética de rendimiento en ciencias naturales con 100 reactivos

Área de contenido	Proceso		
	Conocimiento objetivo*	Suficiencia de la información†	Razonamiento inferencial‡
Astronomía	8	3	3
Botánica	6	7	2
Química	10	5	4
Geología	10	5	2
Física	8	5	6
Zoología	8	5	3
Totales	50	30	20

*Conocimiento objetivo: los reactivos pueden responderse con base en el simple reconocimiento de hechos básicos.

†Suficiencia de la información: los reactivos requieren del empleo de información proporcionada en el texto escrito.

‡Razonamiento inferencial: los reactivos pueden responderse haciendo deducciones u obteniendo conclusiones.

que vayan desde el recuerdo simple hasta el razonamiento inferencial. Al proporcionar una tabla de especificaciones previa a la etapa de redacción de reactivos, el autor de la prueba puede garantizar que el instrumento resultante contenga un equilibrio apropiado en cuanto a cobertura de temas y que detecta un rango deseado de habilidades cognitivas. En el cuadro 4–6 se presenta una tabla de especificaciones hipotética, pero realista.

FORMATOS DE LOS REACTIVOS

En lo que se refiere al método a través del cual se han de evaluar los atributos psicológicos, el autor de pruebas se enfrenta a docenas de opciones. De hecho, sería fácil escribir un capítulo completo sólo sobre este tema. Para una reseña sobre formatos de reactivos, el lector interesado debería consultar a Bausell (1986), Jensen (1980) y Wesman (1971). En esta sección se revisarán de manera rápida las ventajas e inconvenientes de las variedades más comunes de reactivos de una prueba.

En el caso de pruebas de aplicación grupal sobre intelecto y rendimiento, la técnica a elegir es la de preguntas de opción múltiple. Por ejemplo, un reactivo en una prueba de rendimiento sobre historia estadounidense podría incluir esta combinación de planteamiento y opciones:

El presidente de Estados Unidos durante la Guerra Civil fue:

a. Washington.
b. Lincoln.
c. Hamilton.
d. Wilson.

Los defensores de la metodología de opción múltiple afirman que los reactivos construidos de manera apropiada pueden medir tanto el conocimiento conceptual como el objetivo. Las pruebas de opción múltiple también permiten la calificación mecánica rápida y objetiva. Lo que es más, la equidad de las preguntas de opción múltiple se puede probar (¡o en ocasiones refutar!) con los procedimientos muy sencillos de análisis de reactivos que se analizan posteriormente. Las principales desventajas de las preguntas de opción múltiple son, en primer lugar, la dificultad de escribir buenas opciones distractoras y, en segundo, la posibilidad de que la presencia de la respuesta pueda darle un indicio de la contestación correcta a una persona con un conocimiento vago. Las pautas para la redacción de buenos reactivos de opción múltiple se listan en el cuadro 4–7.

Los métodos de correspondencias son populares en el salón de clases, pero sufren de graves desventajas psicométricas. Un ejemplo de correspondencias es:

Utilice las letras que aparecen del lado izquierdo para establecer la correspondencia entre el nombre y la obra alcanzada:

A. Binet _____ tradujo una de las principales pruebas de inteligencia.

Cuadro 4–7. Pautas para la redacción de reactivos de opción múltiple

Elija palabras que tengan significados precisos.
Evite la sintaxis compleja o extraña.
Incluya toda la información necesaria para la selección de la respuesta.
Incluya lo más posible de la pregunta dentro del planteamiento.
No tome los planteamientos al pie de la letra de los libros de texto.
Utilice opciones de igual extensión y fraseo paralelo.
Utilice "ninguna de las anteriores" y "todas las anteriores" en raras ocasiones.
Reduzca al mínimo el empleo de negativos como no.
Evite la utilización de palabras no funcionales.
Evite la especificidad innecesaria en el planteamiento.
Evite indicios innecesarios de la respuesta correcta.
Presente los reactivos a otras personas para escrutinio editorial.

B. Woodworth	_____	ninguna correlación entre calificaciones y pruebas mentales.
C. Cattell	_____	desarrolló un inventario de personalidad de falso-verdadero.
D. McKinley	_____	batería de pruebas sensoriomotoras.
E. Wissler	_____	desarrolló la primera prueba útil de inteligencia.
F. Goddard	_____	prueba de detección del trastorno emocional.

El problema más grave con las correspondencias es que las respuestas no son independientes —en general, si el examinando pierde una correspondencia, esto lo obliga a fallar en otra. Otro problema es que las opciones en una correspondencia deben relacionarse de manera estrecha o la pregunta será demasiado fácil.

Cuando se trata de pruebas aplicadas de manera individual, el procedimiento a elegir es el reactivo objetivo con respuesta corta. De hecho, es frecuente que los tipos más sencillos y directos de preguntas posean la mejor confiabilidad y validez. Un caso específico es la subprueba de Vocabulario del WAIS-III, que consiste sólo de pedirle a la persona evaluada que defina palabras. Ésta tiene una confiabilidad muy elevada (.96) y en general se considera como la mejor medida individual de la inteligencia general en el WAIS-III (Gregory, 1999).

Con frecuencia, las pruebas de personalidad utilizan preguntas de falso-verdadero debido a que son fáciles de comprender para los individuos. A la mayoría de las personas les resulta sencillo responder con verdadero o falso a reactivos como:

V F
____ ____ Me gustan las revistas deportivas.

Los críticos de este enfoque han señalado que las respuestas a dichas preguntas pueden reflejar la deseabilidad social más que los rasgos de personalidad (Edwards, 1961). Un formato alternativo para contrarrestar este problema es la **metodología de elección forzada**, donde la persona evaluada debe elegir entre dos opciones igualmente deseables (o indeseables):

Cuál de estas opciones preferiría hacer:
____ Limpiar un galón de jarabe del piso.
____ Trabajar como voluntario por medio día en un asilo de ancianos.

Aunque el enfoque de elección forzada tiene muchas propiedades psicométricas (Zavala, 1965), los autores de pruebas de personalidad no se han lanzado a elegir esta interesante metodología.

EXAMEN DE REACTIVOS

Los psicómetras saben que numerosos reactivos de prueba de la reserva original del piloteo se descartarán o revisarán a medida que se prosiga con el desarrollo de la prueba. Por tal razón, los autores producen inicialmente muchos

reactivos adicionales, quizás el doble de los que pretenden utilizar. Así que ¿cómo se selecciona la muestra final de reactivos de una prueba a partir de la reserva inicial de éstos? Los autores de pruebas utilizan el análisis de reactivos, una familia de procedimientos estadísticos, para identificar los mejores. En general, el propósito del análisis de reactivos consiste en determinar cuáles de ellos se conservarán, cuáles se revisarán y cuáles se eliminarán. Cuando se realiza un análisis detallado de los reactivos, el autor podría utilizar un índice de dificultad, uno de confiabilidad, uno de validez, una curva reactivo-característica o un índice de discriminación de los reactivos. En este punto hay que volver la atención hacia una breve revisión de estos enfoques estadísticos para el análisis de reactivos. Los lectores que deseen una discusión y crítica a profundidad sobre estos temas deberían consultar a Hambleton (1989) y Nunnally (1978).

ÍNDICE DE DIFICULTAD DEL REACTIVO

La dificultad del reactivo para un solo reactivo de prueba se define como la proporción de personas examinadas en una muestra considerable de piloteo que responden correctamente éste. En cualquier reactivo individual i, el índice de dificultad del reactivo es p_i, que varía de 0.0 a 1.0. Un reactivo con dificultad de .2 es más complicado que uno con una dificultad de .7, debido a que son menos las personas que lo respondieron de manera correcta.

El **índice de dificultad del reactivo** es una herramienta útil para identificar a aquellos que deberían alterarse o descartarse. Supóngase que un reactivo tiene un índice de dificultad cercano a 0.0, lo cual significa que casi todos lo han respondido de manera incorrecta. Por desgracia, este reactivo es improductivo en sentido psicométrico debido a que no proporciona información acerca de las diferencias entre personas examinadas. Para la mayor parte de las aplicaciones, el reactivo debería volver a redactarse o eliminarse. Lo mismo puede decirse de uno con índice de dificultad cercano a 1.0, donde casi todos los participantes dieron la respuesta correcta.

¿Cuál es el nivel óptimo de dificultad del reactivo? En general, las dificultades de reactivo que giran alrededor de .5, con un rango entre .3 y .7, aumentan al máximo la información que la prueba proporciona acerca de las diferencias entre personas evaluadas. No obstante, esta regla empírica está sometida a una salvedad importante y a una excepción muy significativa.

En el caso de reactivos de verdadero-falso o de opción múltiple, el nivel óptimo de dificultad del reactivo necesita ajustarse con relación a los efectos de las conjeturas. En el caso de una prueba de verdadero-falso, se puede obtener un nivel de dificultad de .5 cuando los examinandos simplemente adivinan la respuesta. Así, el nivel óptimo de dificultad de dichos reactivos debería ser de .75 (a la mitad entre .5 y 1.0). En general, el nivel óptimo de dificultad del reactivo se puede calcular con la fórmula $(1.0 + c)/2$ donde c se refiere al nivel de éxito aleatorio. Así, para una prueba de cuatro reactivos de opción múltiple, el nivel de éxito aleatorio es de .25 y el nivel óptimo de dificultad del reactivo sería $(1.0 + .25)/2$ o aproximadamente .63.

Si se ha de utilizar una prueba para seleccionar a un grupo extremo por medio de una puntuación de corte, puede ser deseable la selección de reactivos con niveles de dificultad fuera del rango de .3 a .7. Por ejemplo, una prueba que se emplee para seleccionar a estudiantes de posgrado para una universidad que sólo acepta a unos cuantos elegidos de entre sus muchos solicitantes, debería contener muchos reactivos muy difíciles. Una prueba que se utilice para asignar a niños a un programa de regularización educativa debería contener muchos reactivos sumamente fáciles. En ambos casos, habrá una discriminación útil entre las personas examinadas cerca de la puntuación de corte —una puntuación muy alta para las admisiones al posgrado y muy baja para los estudiantes elegibles para regularización— pero poca discriminación entre los individuos restantes (Allen y Yen, 1979).

ÍNDICE DE LA CONFIABILIDAD DEL REACTIVO

Un autor de pruebas puede desear un instrumento que tenga un alto nivel de consistencia interna, en el que los reactivos sean razonablemente homogéneos. Una manera sencilla de determinar si un reactivo individual "se lleva" con los restantes en la prueba consiste en correlacionar las puntuaciones obtenidas en ese reactivo con aquéllas de la prueba completa. Sin embargo, por lo común los reactivos individuales son correctos o incorrectos (a menudo se les califica con 1 o 0), mientras que las puntuaciones totales constituyen una variable continua. A fin de correlacionar estos dos tipos diferentes de puntuaciones, es necesario utilizar un tipo especial de método estadístico denominado **coeficiente de correlación biserial de puntos**. La fórmula para el cálculo de este coeficiente de correlación es equivalente a la *r* de Pearson que se analizó antes y el coeficiente de correlación biserial de puntos trasmite en mucho el mismo tipo de información concerniente a la relación entre dos variables (una de las cuales casualmente es dicotómica y se califica con 0 o 1). En general, entre más alta sea la correlación biserial de puntos r_{rT} entre un reactivo individual y la puntuación total, más útil será éste desde la perspectiva de la consistencia interna.

La utilidad de un reactivo dicotómico de una prueba también se determina por el grado en que las puntuaciones obtenidas en él se distribuyen entre los dos resultados de 0 y 1. Aunque suena incongruente, es posible calcular la desviación estándar para reactivos dicotómicos; como ocurre con una variable con puntuación continua, la desviación estándar de un reactivo dicotómico indica el grado de dispersión de las puntuaciones. Si un reactivo individual tiene una desviación estándar de cero, todos los individuos están obteniendo la misma puntuación (todos correctos o todos incorrectos). Entre más se acerque el reactivo a una división de 50-50 de puntuaciones correctas e incorrectas, mayor será la desviación estándar. En general, entre más grande sea la desviación estándar de un reactivo, más útil es éste para la escala completa. Aunque no se proporcionará la derivación, se puede demostrar que la desviación estándar de la puntuación del reactivo, s_r, para un reactivo con calificación dicotómica puede calcularse con la fórmula:

$$S_r = \sqrt{p_r(1 - p_i)}$$

Puede resumirse el análisis realizado hasta este momento del siguiente modo: el valor potencial de un reactivo de prueba calificado de manera dicotómica depende conjuntamente de su consistencia interna, como se indica por el índice obtenido de la correlación con la puntuación total (r_{rT}), y también de su variabilidad según el índice obtenido de su desviación estándar (s_r). Si se calcula el producto de estos dos índices, se obtiene $s_r r_{rT}$, que es el **índice de confiabilidad del reactivo**. Considérense las características de un reactivo que posee un índice relativamente grande de confiabilidad. Éste exhibiría fuerte consistencia interna y produciría una buena dispersión de puntuaciones entre sus dos alternativas. El valor de este índice en la construcción de pruebas es, dicho de manera sencilla, el siguiente: al calcular el índice de confiabilidad de todos los reactivos en la prueba preliminar, pueden eliminarse los "más externos" que tienen el menor valor en este índice. Dichos reactivos poseerían una consistencia interna muy baja o dispersión de puntuaciones débil y, por tanto, no contribuirían a los objetivos de la medición.

ÍNDICE DE LA VALIDEZ DEL REACTIVO

En muchas aplicaciones, es importante que una prueba posea una validez concurrente o predictiva más elevada posible. En estos casos, una pregunta preponderante que gobierna la construcción de pruebas es: ¿cuánto contribuye cada reactivo preliminar de una prueba a la predicción precisa del criterio? El **índice de validez del reactivo** es una herramienta útil en la búsqueda de los psicómetras por identificar reactivos de una prueba que sean útiles en términos

predictivos. Mediante el cálculo del índice de validez de cada reactivo en la prueba preliminar, el autor puede identificar aquellos que son ineficaces, eliminarlos o volver a redactarlos, y producir un instrumento revisado que tenga mayor utilidad práctica.

El primer paso para establecer el índice de validez de un reactivo consiste en calcular la correlación biserial de puntos entre la puntuación del reactivo y aquélla en la variable de criterio. En general, entre más alta sea la correlación biserial de puntos r_{rC} entre las puntuaciones del reactivo individual y del criterio, más útil será el reactivo desde el punto de vista de la validez predictiva. Como se señaló antes, la utilidad de un reactivo también depende de su desviación estándar s_r. Así, el índice de validez del reactivo consiste del producto entre la desviación estándar y la correlación biserial de puntos: $s_r r_{rC}$.

CURVAS REACTIVO-CARACTERÍSTICA

Una **curva reactivo-característica** (CRC) es una representación gráfica de la relación entre la probabilidad de una respuesta correcta y la posición de la persona evaluada en el rasgo subyacente que mide la prueba. No obstante, no se tiene acceso directo a los rasgos subyacentes, de modo que deben utilizarse las puntuaciones observadas para estimar las cantidades del rasgo.

Se traza una CRC independiente para cada reactivo, con base en una gráfica de las puntuaciones totales de una prueba en el eje horizontal contra la proporción de personas que aprueban el reactivo en el eje vertical (figura 4–8). De hecho, una CRC es una idealización matemática de la relación entre la probabilidad de una respuesta correcta y la cantidad del rasgo que poseen aquellos que responden a la prueba. Diferentes modelos de CRC utilizan diversas funciones matemáticas con base en las suposiciones iniciales. El modelo más sencillo de CRC es el de Rasch, que se basa en la teoría de respuesta al reactivo del matemático danés Georg Rasch (1966). El modelo de Rasch es el más sencillo porque realiza sólo dos suposiciones: 1) los reactivos de prueba son unidimensionales y miden un rasgo común y 2) los reactivos de prueba varían en un continuo de nivel de dificultad.

En general, un buen reactivo tiene una CRC con una pendiente positiva. Si la capacidad para resolver un reactivo en particular se distribuye de manera normal, la CRC se asemejará a una **ojiva normal** (curva **a** en la figura 4–8). Ésta es, simplemente, la distribución normal trazada en forma acumulativa.

La forma deseada de la CRC depende del propósito de la prueba. Los puristas de la psicometría preferirían que la CRC del reactivo de una prueba se aproximara a la ojiva normal, porque esta curva es conveniente para hacer deducciones matemáticas acerca del rasgo subyacente (Lord y Novick, 1968). Sin embargo, para decisiones de selección que se basan en puntuaciones de corte, se prefiere una función escalonada. Por ejemplo, cuando se combina con otros reactivos similares, el reactivo que produjo la curva **b** en la figura 4–8 sería el mejor para seleccionar individuos con altos niveles del rasgo medido.

Las CRC son especialmente útiles para identificar reactivos que tienen un desempeño diferente en los subgrupos de personas examinadas (Allen y Yen, 1979). Por ejemplo, un autor de pruebas puede descubrir que un reactivo tiene un desempeño diferente en varones y mujeres. Ése podría ser el caso de una pregunta parcial en cuanto a género que implique

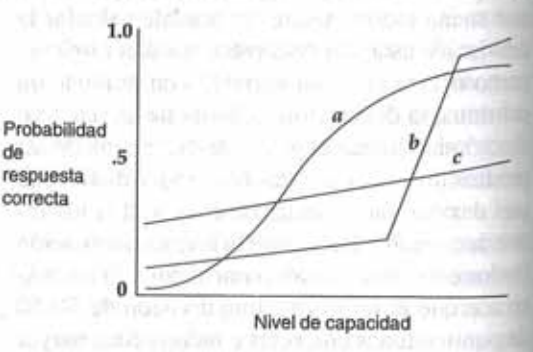


Figura 4–8. Algunos ejemplos de curvas reactivo-característica.

datos sobre fútbol. Para los varones, la CRC en este reactivo podría tener la pendiente positiva deseada, mientras que para mujeres ésta podría ser plana (como la curva **c** en la figura 4–8). Los reactivos con CRC que difieren entre subgrupos de personas evaluadas deben revisarse o eliminarse.

La teoría subyacente de la CRC también se conoce como **teoría de respuesta al reactivo** y **teoría de rasgo latente**. Nunnally (1978) ha cuestionado la utilidad de este enfoque al señalar que la suposición de unidimensionalidad de la prueba (implícita en la CRC que representa gráficamente el porcentaje que aprueba contra el eje horizontal unidimensional del valor del rasgo) se viola cuando se consideran muchas pruebas psicológicas. Si no hubiera serios problemas técnicos y prácticos, "uno se pregunta por qué no se ha adoptado desde hace largo tiempo la teoría de la CRC para la construcción y calificación de pruebas" (Nunnally, 1978).

Los méritos del enfoque de la CRC todavía se discuten. La teoría de la CRC parece particularmente apropiada para ciertos tipos de pruebas adaptativas por computadora (PAC) en las que la persona responde a un conjunto individualizado y único de reactivos que se califican entonces en una escala uniforme subyacente (Weiss, 1983). El enfoque de las PAC para la evaluación no sería posible en ausencia de los métodos de CRC para la medición. Las PAC se analizan en el tema 15A, Evaluación por computadora y el futuro de las pruebas. Los lectores que deseen un análisis más detallado de la CRC y de otros modelos de rasgo latente deberían consultar a Anastasi (1988), Hambleton (1989) y Wright y Stone (1979).

ÍNDICE DE DISCRIMINACIÓN DEL REACTIVO

A partir del análisis de las CRC, debería quedar claro que un reactivo de prueba eficaz es aquel que discrimina entre los individuos con puntuaciones altas y bajas en la prueba completa. Un reactivo ideal de una prueba es el que aprueba a la mayoría de las personas con altas puntuaciones y falla con la mayoría de los que tienen bajas puntuaciones (véase curva **a** en la figura 4–8). La simple inspección visual de la CRC proporciona una base rudimentaria para estimar la capacidad de discriminación de un reactivo de una prueba: si la pendiente de la curva es positiva y, de preferencia, tiene forma de ojiva, el reactivo está haciendo un buen trabajo en cuanto a separar a las personas con puntuaciones altas y bajas. Pero la inspección visual no es un procedimiento totalmente objetivo; lo que se requiere es una herramienta estadística que resuma el poder de discriminación de los reactivos individuales de una prueba.

Un **índice de discriminación del reactivo** es un índice estadístico de la eficiencia con la que un reactivo discrimina entre personas que obtienen altas y bajas puntuaciones en la prueba completa. Existen muchos índices de discriminación del reactivo, que incluyen medidas indirectas como la r_{rT}, la correlación biserial de puntos entre las puntuaciones en un reactivo individual y la puntuación total en la prueba. No obstante, aquí se restringirá el examen a una medida directa, el índice de discriminación del reactivo, que se simboliza por la letra minúscula *d*, escrita en cursivas. En una base de reactivo por reactivo, este índice compara el desempeño de los participantes en las regiones superior e inferior de la puntuación total de prueba. En general, los rangos superior e inferior se definen como el 10 a 33% de personas dentro de la muestra con puntuaciones más altas y más bajas. Si las puntuaciones totales de la prueba tienen una distribución normal, la comparación óptima es entre 27%, de personas con puntuaciones más altas y 27% con puntuaciones más bajas. Si la distribución de las puntuaciones totales de una prueba es más plana que la curva normal, el porcentaje óptimo es más grande, cercano a 33%. En el caso de la mayor parte de las aplicaciones, cualquier porcentaje entre 25 y 33% producirá estimados similares de *d* (Allen y Yen, 1979).

El índice de discriminación del reactivo para un reactivo de prueba se calcula con la fórmula:

$$d = (S–I)/N$$

donde *S* es el número de personas evaluadas que se encuentran en el rango superior y que respondieron de manera correcta al reactivo, *I* es el número de personas evaluadas que están del rango inferior y que respondieron al reactivo en forma correcta y *N* es el número total de personas evaluadas en los rangos superior o inferior.

Ilustre el cálculo y empleo de *d* con un ejemplo hipotético. Supóngase que un autor de pruebas ha construido una versión preliminar de una prueba de rendimiento de opción múltiple y ha aplicado el examen a una muestra de piloteo de 400 alumnos de secundaria. Después de calcular las puntuaciones totales de cada individuo, el autor identifica a 25% con puntuaciones más altas y a 25% con puntuaciones más bajas dentro de la muestra. Dado que hay 100 alumnos en cada grupo (25% de 400), en la fórmula anterior *N* será 100. A continuación, el autor determina el número de estudiantes en los rangos superior e inferior que respondieron a cada reactivo de manera correcta. El cálculo de *d* para cada uno se reduce simplemente a sustituir estos valores dentro de la fórmula *(S–I)/N*. Por ejemplo, supóngase que en el primer reactivo 49 alumnos en el rango superior respondieron de manera correcta, mientras que 23 del rango inferior lo hicieron correctamente. En este reactivo, *d* es igual a (49–23)/100 o .26.

De la fórmula para *d* es evidente que este índice puede variar de –1.0 a +1.0. Véase también que un valor negativo en *d* es señal de alarma de que un reactivo de prueba necesita revisión o sustitución. Después de todo, dicho resultado indicaría que fueron más los individuos con bajas puntuaciones quienes respondieron de manera correcta al reactivo que aquéllos con altas puntuaciones. Si *d* es igual a cero, hay un número exactamente igual de individuos con bajas y altas puntuaciones que respondieron de manera correcta al reactivo; dado que éste no discrimina en absoluto entre personas con altas y bajas puntuaciones, entonces debería revisarse o eliminarse. Es preferible un valor positivo para *d* y entre más se acerque a +1.0, mejor. El cuadro 4–8 ilustra los índices de discriminación de seis reactivos de la prueba hipotética que se propone aquí.

Un autor de pruebas puede complementar el enfoque de discriminación de los reactivos al inspeccionar el número de personas evaluadas en los grupos con puntuaciones más altas y más bajas que eligieron cada una de las alternativas incorrectas. Si un reactivo de opción múltiple está bien redactado, las alternativas incorrectas deben ser igualmente atractivas para los participantes que no conocen la respuesta correcta. Por supuesto, se esperaría que los individuos con puntuaciones altas eligieran la alternativa correcta con mayor frecuencia que aquéllos con bajas puntuaciones —ese es el propósito de calcular índices de discriminación del reactivo. Pero, además, un buen reactivo debería mostrar una dispersión proporcional de opciones incorrectas tanto para los individuos con altas puntuaciones como para aquéllos con bajas puntuaciones.

Supóngase que se investigan las opciones tomadas por 100 individuos con altas puntuaciones y 100 con bajas puntuaciones en una

Cuadro 4–8. Índices de discriminación del reactivo de seis reactivos hipotéticos

Reactivo	S	I	(S–1)/N	Comentario
1	49	23	.26	Reactivo muy bueno con dificultad elevada.
2	79	19	.60	Excelente reactivo, pero se obtiene en raras ocasiones.
3	52	52	.00	Reactivo deficiente que debería revisarse.
4	100	0	1.00	Reactivo ideal, pero nunca se obtiene.
5	20	80	–.60	Reactivo muy inadecuado que debería eliminarse.
6	0	100	–1.00	El peor reactivo posible en términos teóricos.

prueba hipotética de opción múltiple. Las opciones correctas se indican con un asterisco (*). El reactivo 1 demuestra el patrón deseado de respuestas, con una dispersión aproximadamente igual de las opciones incorrectas.

	Alternativas				
Reactivo 1	*a*	*b*	*c**	*d*	*e*
Altas puntuaciones	5	6	80	5	4
Bajas puntuaciones	15	14	40	16	15

En el reactivo 2, se puede ver que ninguno de los alumnos examinados eligió la alternativa *d*. Ésta debe reemplazarse con un medio de distracción más atractivo:

Reactivo 2	*a*	*b**	*c*	*d*	*e*
Altas puntuaciones	5	75	10	0	10
Bajas puntuaciones	21	34	20	0	25

El reactivo 3 probablemente es deficiente, a pesar del hecho de que discrimina entre los participantes con altas y bajas puntuaciones. El problema obvio es que los alumnos evaluados que tienen altas puntuaciones prefieren la alternativa *a* en lugar de la alternativa correcta, *d*:

Reactivo 3	*a*	*b*	*c*	*d**	*e*
Altas puntuaciones	43	6	5	37	9
Bajas puntuaciones	20	19	22	10	29

Quizás si se vuelve a escribir la alternativa *a*, este reactivo pudiera rescatarse. En cualquier caso, el asunto principal aquí es que los autores de pruebas deben escudriñar en cada rincón de cada reactivo de prueba a través de todos los medios posibles, incluyendo la inspección visual del patrón de respuestas.

REITERACIÓN: LOS MEJORES REACTIVOS

De todos los métodos de análisis de reactivos que se presentan antes, ¿cuáles debe utilizar el autor de pruebas para identificar los mejores reactivos para una prueba? La respuesta a esta pregunta no es ni simple ni sencilla. Después de todo, la elección del "mejor" reactivo depende de los objetivos del autor de la prueba. Por ejemplo, un psicólogo inclinado teóricamente hacia la investigación podría querer un instrumento de medición con la mayor consistencia interna posible; los índices de confiabilidad del reactivo son cruciales para este objetivo. Un administrador universitario con mentalidad práctica podría desear un instrumento con la mayor validez de criterio posible; los índices de validez del reactivo serían útiles para este propósito. Un especialista en retraso mental, orientado a la terapéutica, podría necesitar una prueba de inteligencia con un efecto mínimo de límite inferior; los índices de dificultad del reactivo serían útiles en este sentido. En resumen, no existe un solo método preferido para la selección de reactivos que sea idealmente adecuado para todos los contextos de evaluación y desarrollo de pruebas.

REVISIÓN DE LA PRUEBA

El propósito del análisis de reactivos que se analizó anteriormente consiste en identificar los reactivos improductivos de la prueba preliminar de modo que se puedan revisar, eliminar o reemplazar. Muy pocas pruebas surgen incólumes de este proceso. Es común dentro del proceso evolutivo del desarrollo de pruebas que se descarten muchos reactivos, que otros se refinen y que se añadan nuevos. La repercusión inicial es que surge una prueba nueva y ligeramente diferente. Es probable que esta prueba revisada contenga más reactivos discriminativos, con mayor confiabilidad y precisión predictiva más elevada —pero se sabe que estas mejorías son verdaderas sólo en el caso de la primera muestra del piloteo.

El siguiente paso en el desarrollo de una prueba consiste en obtener nuevos datos de una segunda muestra de piloteo. Por supuesto, estos individuos evaluados deben ser similares a aquéllos a los que, en última instancia, se dirige la prueba. El propósito de reunir datos adicionales de una prueba es repetir de nuevo los

procedimientos de análisis de reactivos. Si los cambios posteriores son de una variedad menor de tipo afinación, el autor puede decidir que la prueba es satisfactoria y que está lista para un estudio de validación cruzada, el cual se analiza en la siguiente sección. Si se requieren cambios mayores, es deseable recolectar datos de una tercera e, incluso, quizá de una cuarta muestra de piloteo. Pero en algún punto deben cesar estos pequeños ajustes psicométricos; el autor debe proponer un instrumento terminado y proseguir con el siguiente paso, la validación cruzada.

VALIDACIÓN CRUZADA

Cuando se utiliza una muestra de piloteo para determinar que una prueba posee validez relacionada con el criterio, la evidencia es bastante preliminar y tentativa. En el desarrollo de pruebas, es una práctica prudente la búsqueda de confirmación fresca e independiente de la validez de la prueba antes de proceder con la publicación. El término **validación cruzada** se refiere a la práctica de utilizar la ecuación original de regresión en una muestra nueva para determinar si la prueba predice el criterio de la misma manera que en que lo hace la original. Ghiselli, Campbell y Zedeck (1981) resumen la fundamentación para la validación cruzada:

> Ya sea que los reactivos se elijan con base en concordancia empírica o que se corrijan o ponderen, deberían considerarse como específicos de la muestra que se empleó para los análisis estadísticos, a menos que se obtengan datos adicionales. Esto es necesario porque existe la probabilidad de que los resultados obtenidos se hayan beneficiado de factores aleatorios que operaron en ese grupo y, por tanto, son aplicables sólo para la muestra estudiada.

REDUCCIÓN DE LA VALIDEZ

Un descubrimiento común en la investigación de validación cruzada es que una prueba predice el criterio relacionado de manera menos precisa con la nueva muestra de examinandos que con la original de piloteo. El término **reducción de la validez** se aplica a este fenómeno. Por ejemplo, un medio de predicción biográficamente basado del potencial de ventas podría tener un desempeño bastante adecuado con la muestra de individuos que se empleó para desarrollar el instrumento, pero demuestra menos validez cuando se aplica a un nuevo grupo de personas evaluadas. Mitchell y Klimoski (1986) estudiaron la reducción de la validez de un instrumento diseñado para predecir cuáles estudiantes tendrán éxito en bienes raíces, al medirlo por el criterio real de obtener una licencia en bienes raíces dos años después. En un análisis basado en la muestra utilizada para derivar la prueba, la prueba predictiva con base biográfica se correlacionó en .6 con el criterio. Pero cuando se puso a prueba este mismo instrumento con una nueva muestra de estudiantes de bienes raíces, la correlación con el criterio fue menor, cercana a .4; lo cual demuestra una reducción típica de la validez.

La reducción de la validez es una parte inevitable del desarrollo de pruebas y subraya la necesidad de obtener una validación cruzada. En la mayor parte de los casos, la reducción es leve y el instrumento sobrevive al desafío de la validación cruzada. No obstante, la reducción en la validez de una prueba puede representar un grave problema cuando las muestras de derivación y de validación cruzada son pequeñas el número de reactivos potenciales de una prueba es grande y éstos se eligen con una base puramente empírica, sin fundamentación teórica.

Un artículo clásico escrito por Cureton (1950) demuestra el peor escenario posible: utilizar una muestra muy pequeña para seleccionar reactivos con concordancia empírica a partir de una reserva de reactivos grande y después validar la prueba con la misma muestra. El criterio en su estudio era el promedio de calificaciones, dividido de manera artificial en las calificaciones de B o mejores y calificaciones menores a B. Sus reactivos de "prueba" consistían de 85 etiquetas, numeradas de un lado. Con cada uno de los 29 estudiantes, éstas se revolvían en un recipiente y se dejaban caer sobre una mesa. Se registraban todas las que

caían con los números hacia arriba como indicación de la presencia de ese "reactivo" para el estudiante. A continuación, Cureton llevó a cabo un análisis de reactivos utilizando las calificaciones divididas como criterio. Con base en este análisis, se encontró que 24 reactivos tenían la capacidad máxima de predicción de las calificaciones de los estudiantes. Nueve de ellos se presentaron con mayor frecuencia entre los estudiantes con mayores calificaciones y éstos se ponderaron como +1. Quince reactivos ocurrieron con mayor frecuencia en el caso de los estudiantes con bajas calificaciones y éstos se ponderaron como –1. La puntuación de esta prueba (irónicamente llamada "Prueba Proyectiva B de Psicoquinesia") consistió de la suma de estas 24 ponderaciones de reactivo.

A pesar de la naturaleza insensata de su prueba, Cureton (1950) encontró que las puntuaciones de la prueba se correlacionaban en .82 con las calificaciones. Por supuesto, la fortaleza de esta correlación se debía por completo a que se obtenía beneficio del azar. Si se realizara una serie de estudios de validación cruzada con nuevas muestras de estudiantes, es probable que la correlación entre la Prueba Proyectiva B de Psicoquinesia y las calificaciones giraría más o menos alrededor de cero, porque ésta carece por completo de validez predictiva. Esto constituye una importante lección que se aplica también a las pruebas serias: debe demostrarse la validez a través de validación cruzada, no debe suponerse que existe simplemente con base en las solemnes intenciones de un nuevo instrumento.

REALIMENTACIÓN DE LAS PERSONAS EVALUADAS

En la revisión de la prueba, la realimentación de las personas evaluadas es una fuente potencialmente valiosa de información que en general pasan por alto los autores de pruebas. Se puede ilustrar este enfoque con la investigación de Nevo (1992). Este autor desarrolló el *Examinee Feedback Questionnaire* (Cuestionario de Realimentación del Examinando; EFeQ) para estudiar el *Inter-University Psychometric Entrance Examination* (Examen Psicométrico de Ingreso Interuniversitario), uno de los principales requisitos para admisión a las seis universidades de Israel. El Examen de Ingreso Interuniversitario es una prueba grupal que consiste de cinco subpruebas de opción múltiple: Conocimientos generales, Razonamiento figurativo, Comprensión, Razonamiento matemático e Inglés. El EFeQ se diseñó como una posprueba anónima que se aplica de inmediato después de contestar el Examen de Ingreso Interuniversitario.

El EFeQ es un cuestionario corto y sencillo, diseñado para obtener las opiniones sinceras de las personas evaluadas en cuanto a los siguientes rasgos de la matriz prueba-examinador-persona evaluada:

- Conducta de los examinadores
- Condiciones de prueba
- Claridad de las instrucciones de examen
- Conveniencia en el uso de la hoja de respuestas
- Percepción sobre la idoneidad de la prueba
- Percepción sobre imparcialidad cultural de la prueba
- Percepción acerca de la suficiencia de tiempo
- Percepción acerca de la dificultad de la prueba
- Respuesta emocional ante la prueba
- Nivel de conjeturas
- Trampa de parte de la persona evaluada u otros.

La pregunta final en el EFeQ es un ensayo libre: "Estamos interesados en cualesquiera comentarios o sugerencias que pueda tener para mejorar el examen". En la figura 4–9 se proporcionan algunos ejemplos de las preguntas de realimentación similares a las del EFeQ.

Nevo (1992) determinó que el cuestionario EFeQ posee una confiabilidad modesta, con una confiabilidad *test-retest* de aproximadamente .70. Aparte de las propiedades psicométricas de su escala, la tradición de pedirles realimentación a los examinandos acerca de las pruebas ha resultado invaluable. El examen de ingreso interuniversitario se modificó de diversas maneras en respuesta a esta realimentación: el

¿Cuál es su opinión acerca de la cantidad de tiempo que se concede a cada prueba? Marque cada casilla con un número del 1 al 5 según las siguientes calificaciones:

5	4	3	2	1
Demasiado tiempo	Mucho tiempo	Tiempo adecuado	Poco tiempo	Muy poco tiempo

☐ Conocimientos generales.
☐ Razonamiento figurativo.
☐ Comprensión.
☐ Razonamiento matemático.
☐ Inglés.

¿Usted u otra persona hizo trampa en este examen? Por favor marque las casillas según se aplique. Puede marcar más de una casilla.

☐ Sí—obtuve una copia de la prueba.
☐ Sí—uno de los examinadores me ayudó de manera ilegal.
☐ Sí—uno de los examinadores me ayudó durante la prueba.
☐ Sí—yo ayudé a otro de los examinandos.
☐ Sí—utilicé notas ocultas durante la prueba.
☐ Sí—vi a otra persona haciendo trampa.
☐ No—no hice ninguna trampa.
☐ No—no vi que nadie más hiciera trampa.

Figura 4–9. Ejemplos de reactivos del cuestionario de realimentación del examinando.
Fuente: Basado en Nevo, B. (1992). Examinee feedback: Practical guidelines. In M. Zeidner and R. Most (Eds.), *Psychological testing: An inside view.* Palo Alto, Calif.: Consulting Psychologists Press.

mato de hoja de respuestas se cambió en los sentidos sugeridos por las personas evaluadas; el límite de tiempo se amplió para pruebas específicas de las que se informó que eran demasiado rápidas, y se eliminaron ciertos reactivos percibidos como parciales o poco equitativos. Además, se revisaron y estrecharon las medidas de seguridad a fin de reducir el copiado, que era más predominante de lo que los examinadores habían anticipado. Nevo (1992) también cita una ventaja oculta de los cuestionarios de realimentación: transmiten el mensaje de que existe alguien a quien le interesa escuchar, lo cual reduce el estrés posterior al examen. Los cuestionarios de realimentación de los examinandos deben volverse una práctica rutinaria en las pruebas grupales estandarizadas.

PUBLICACIÓN DE LA PRUEBA

El proceso de construcción de pruebas no termina con la obtención de los datos de validación cruzada. El autor de pruebas también debe supervisar la producción de los materiales de la prueba, publicar un manual técnico y producir otro del usuario. Se pueden ofrecer varias pautas relevantes para cada uno de estos pasos finales, como se describe en las siguientes secciones. Por último, se cierra este capítulo con un provocativo comentario acerca del conservadurismo de los editores actuales de pruebas.

PRODUCCIÓN DE LOS MATERIALES DE PRUEBA

Los materiales de prueba deben ser amigables para el usuario, si es que han de recibir amplia aceptación de parte de psicólogos y educadores. Así, una pauta inicial para la producción de la prueba es que la presentación física de los materiales debe permitir la aplicación rápida y sin complicaciones. Considérese el reto que representan algunas pruebas de desempeño, donde el examinador debe luchar con el lápiz, la tablilla para anotaciones, el protocolo de prueba, cronómetro, manual de prueba, pantalla para los reactivos, caja del reactivo y un objeto desarmado de cartón, al mismo tiempo que mantiene una conversación con el examinando.

Si el autor de la prueba está en posibilidad de simplificar los deberes del examinador al mismo tiempo que permite que las demandas de la tarea sean iguales para el examinando, el instrumento resultante tendrá una aceptabilidad mucho mayor para los usuarios potenciales. Por ejemplo, si las instrucciones de aplicación pueden resumirse en el protocolo de prueba, el examinador puede hacer a un lado el manual mientras que coloca la tarea frente al examinando. Otra adición bien recibida para la presentación de la prueba psicológica es la encuadernación con engargolado, la cual tiene la posibilidad de colocarse de manera vertical y que muestra las preguntas de prueba en la cara que se presenta al examinando a la vez que proporciona las instrucciones de aplicación en el lado contrario que da hacia el examinador.

MANUAL TÉCNICO Y MANUAL DEL USUARIO

En general, los datos técnicos acerca de un nuevo instrumento se resumen con las referencias apropiadas en un **manual técnico**. En éste, el futuro usuario puede encontrar información acerca del análisis de los reactivos, confiabilidades de escala, estudios de validación cruzada y similares. En algunos casos, esta información se incorpora en el **manual del usuario**, que da instrucciones para la aplicación y también proporciona las pautas para la interpretación de la prueba.

Los manuales de prueba deben comunicar información a muchos grupos diferentes, que abarcan en cuanto a antecedentes y entrenamiento desde los especialistas en medición hasta el maestro en el salón de clases. Los manuales de prueba satisfacen muchos propósitos, como se describe en los *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1985, 1999). Este influyente compendio de los *Standards* sugiere que los manuales de prueba deben alcanzar los siguientes objetivos:

- Describir la fundamentación y usos recomendados de la prueba.
- Proporcionar advertencias específicas en contra del mal uso anticipado de una prueba.
- Citar estudios representativos concernientes a usos generales y específicos de una prueba.
- Identificar los requisitos especiales necesarios para aplicar e interpretar la prueba.
- Proporcionar revisiones, enmiendas y complementos según sea necesario.
- Utilizar material promocional que sea preciso y basado en investigación.
- Citar relaciones cuantitativas entre las puntuaciones de una prueba y los criterios.
- Informar del grado en que se pueden intercambiar los modos alternativos de respuesta (p. ej., folleto contra una hoja de respuesta).
- Proporcionar auxiliares interpretativos apropiados para la persona que realiza la prueba.
- Suministrar evidencia de la validez de cualesquier interpretaciones automáticas de una prueba.

Por último, los manuales de prueba deben proporcionar datos esenciales acerca de la confiabilidad y validez, en vez de referir al usuario hacia otras fuentes —una práctica desafortunada que se encuentra en algunos manuales de prueba.

LAS PRUEBAS SON UN GRAN NEGOCIO

En este momento, el lector deberá advertir la desafiante tarea que enfrenta todo aquel que se propone desarrollar y publicar una nueva prueba. Además de las titánicas proporciones de la empresa, el desarrollo de pruebas es extraordinariamente costoso, lo cual significa que los editores tienen una actitud de naturaleza conservadora en cuanto a la presentación de nuevas pruebas. Jensen (1980) proporciona la siguiente perspectiva tentadora acerca de este tema:

> Producir una nueva prueba de inteligencia general que represente una mejoría verdaderamente significativa, por encima de los instrumentos existentes, sería un proyecto multimillonario que

requeriría gran cantidad de personal constituido de expertos en construcción de pruebas que trabajen durante varios años. Hoy en día poseemos la tecnología psicométrica necesaria para producir pruebas considerablemente mejores que las que se utilizan a nivel popular en la actualidad. Los principales obstáculos son las leyes de derechos de autor, los intereses creados que tienen los editores en las pruebas establecidas en las que ya han realizado enormes inversiones, y la economía de mercado de las pruebas. La mejoría importante de las pruebas no es, de inicio, una aventura comercial atractiva y es probable que tuviera que depender de subsidios a gran escala y a largo plazo de instancias gubernamentales y fundaciones privadas.

RESUMEN

1. La construcción de pruebas consiste de seis etapas entrelazadas: definición de la prueba, selección de un método de escalamiento, construcción y análisis de los reactivos, así como la revisión y publicación de la prueba.
2. Los autores de pruebas necesitan seleccionar un método de escalamiento que se adecue en forma óptima a la manera en que han concebido el rasgo o rasgos que mide la prueba. El concepto de niveles de medición es sumamente relevante en este contexto.
3. Se reconocen cuatro niveles de medición: las escalas nominales constituyen la simple denominación o categorización; las escalas ordinales permiten una clasificación; las escalas de intervalo poseen intervalos iguales; las escalas de razón incorporan todas las características anteriores y también introducen un cero absoluto.
4. Existen docenas de métodos de escalamiento. Ejemplos representativos incluyen el método de escalamiento absoluto, en el que la dificultad del reactivo se localiza en un eje o línea base y se mide en unidades de desviación estándar en un grupo base; las escalas Likert, que presentan reactivos con cinco respuestas ordenadas en un continuo de acuerdo/desacuerdo; y el enfoque de escalamiento racional en el que los reactivos derivados de manera racional se correlacionan con las puntuaciones totales de una prueba.
5. La construcción de los reactivos de una prueba es un procedimiento laborioso y lento. Los autores de pruebas deben evitar los efectos de límite superior e inferior. En un efecto de límite superior, cantidades importantes de personas evaluadas obtienen puntuaciones perfectas o casi perfectas. En un efecto de límite inferior, cantidades significativas de personas obtienen puntuaciones en o cerca de la parte más baja de la escala.
6. Una tabla de especificaciones enumera la información y tareas cognitivas en las que se ha de evaluar a las personas examinadas. En el caso de las pruebas de rendimiento y capacidad, los redactores de los reactivos trabajan, por lo general, a partir de la tabla de especificaciones para garantizar que el instrumento en progreso detecta la mezcla deseada de procesos cognitivos y contenidos de un reactivo.
7. Los reactivos de una prueba pueden redactarse en muchos formatos diferentes, incluyendo opción múltiple, respuestas abiertas, falso-verdadero y elección forzada. Las correspondencias, tan populares en las pruebas para salón de clase, son cuestionables en términos psicométricos, dado que las opciones no son independientes entre sí.
8. El propósito del análisis de los reactivos consiste en determinar cuáles de los iniciales han de conservarse, cuáles revisarse y cuáles eliminarse. Se tienen disponibles muchos procedimientos estadísticos para el análisis de reactivos, que incluyen el índice de dificultad del reactivo, el índice de confiabilidad del reactivo, el índice de validez del reactivo, la curva reactivo-característica y el índice de discriminación del reactivo.

9. El término **validación cruzada** se refiere a la práctica de volver a validar una prueba con una nueva muestra de personas examinadas. **Reducción de la validez** se refiere al fenómeno común en el que una prueba predice el criterio relacionado de manera menos precisa con una nueva muestra que con aquella original de piloteo.

10. Las pruebas deben ser amigables para el usuario si es que han de recibir amplia aceptación de parte de psicólogos y educadores. Por ejemplo, la encuadernación con engargolado, que permite la colocación vertical y que muestra las instrucciones en un lado y los estímulos de prueba en el otro, es especialmente deseable.

TÉRMINOS Y CONCEPTOS CLAVE

Escalas nominales p. 145
Escalas ordinales p. 145
Escala de intervalo p. 145
Escala de razón p. 146
Clasificaciones de expertos p. 147
Método de intervalos de aparición balanceada p. 148
Método de escalamiento absoluto p. 149
Escalas Likert p. 150
Escalas Guttman p. 150
Método de concordancia empírica p. 150
Método de escalamiento racional p. 151
Tabla de especificaciones p. 153
Metodología de elección forzada p. 155
Índice de dificultad del reactivo p. 156
Índice de la confiabilidad del reactivo p. 157
Índice de la validez del reactivo p. 157
Curva reactivo-característica p. 158
Ojiva normal p. 158
Índice de discriminación del reactivo p. 159
Validación cruzada p. 162
Reducción de la validez p. 162
Manual técnico p. 165
Manual del usuario p. 165

CAPÍTULO 5

Pruebas de inteligencia I: Teorías y evaluación preescolar

TEMA 5A
Teorías y medición de la inteligencia

Definiciones de inteligencia

Ejemplo de caso 5–1. Aprendizaje y adaptación como funciones elementales de la inteligencia

Teorías de la inteligencia

Resumen

Términos y conceptos clave

Este capítulo inicia una amplia discusión de las pruebas de inteligencia, un tema tan importante y extenso que se le dedican los siguientes dos capítulos. A fin de comprender las pruebas actuales de inteligencia, el lector necesitará asimilar ciertas definiciones, teorías y prácticas convencionales de evaluación. El objetivo del Tema 5A, Teorías y medición de la inteligencia, consiste en investigar los diversos significados dados al término **inteligencia** y analizar la manera en que las definiciones y teorías han influido en la estructura y contenido de las pruebas de inteligencia. Una justificación importante para este tema es que la comprensión de las teorías de la inteligencia es importante para establecer la validez de constructo de las medidas de CI. En el Tema 5B, Evaluación de las capacidades de lactantes y preescolares, se revisa la naturaleza y aplicación de los principales instrumentos de evaluación de lactantes y después se investiga una cuestión fundamental: la utilidad práctica de

estos instrumentos. Se iniciará con una revisión de las teorías iniciales, tradicionales y contemporáneas acerca de la inteligencia.

La inteligencia es uno de los temas que más se han investigado en psicología. Cada año se publican miles de artículos de investigación sobre su naturaleza y medición. Han surgido revistas como *Intelligence* y *The Journal of Psychoeducational Assessment*, en respuesta al interés académico en esta materia. A pesar de la extensa bibliografía de investigación, la definición de inteligencia aún es esquiva, envuelta en la controversia y el misterio. De hecho, el análisis que se presenta a continuación ilustrará una de las principales paradojas de la evaluación moderna con pruebas: ¡los psicómetras explican mejor la medición de la inteligencia que su definición!

Aunque esta definición ha resultado una labor frustrante, puede obtenerse mucho de la revisión de los esfuerzos antiguos y actuales para aclarar su significado; después de todo, las pruebas de inteligencia no se materializaron de la nada; la mayor parte se fundamentan en una teoría específica de la inteligencia y muchos de los que las elaboran ofrecen una definición del constructo como punto inicial de sus esfuerzos. Por estas razones, se puede comprender y evaluar mejor el carácter multifacético de las pruebas contemporáneas si se revisan primero las principales definiciones y teorías de la inteligencia.

DEFINICIONES DE INTELIGENCIA

Antes de analizar las definiciones de inteligencia es necesario aclarar la naturaleza de la definición en sí. Sternberg (1986) distingue entre **definiciones operacionales** y "reales", lo cual es importante en este contexto. La primera define un concepto en términos de cómo se mide dicho concepto. Boring (1923) llevó este punto de vista a su extremo cuando definió la inteligencia como "aquello que miden las pruebas". Por increíble que parezca, ésta fue una propuesta seria, diseñada en gran medida para evitar los grandes desacuerdos acerca de la definición de inteligencia.

Las definiciones operacionales de inteligencia sufren de dos peligrosas desventajas (Sternberg, 1986). Primero, son circulares. Las pruebas de inteligencia se inventaron para medir la inteligencia, no para definirla. Quienes diseñaron las pruebas nunca tuvieron la intención de que sus instrumentos definieran a la inteligencia. En segundo lugar, las definiciones operacionales impiden un mayor progreso en cuanto a la comprensión de la naturaleza de la inteligencia, porque anulan el análisis acerca de la idoneidad de sus teorías.

Este segundo problema —los efectos potencialmente entorpecedores de depender de las definiciones operacionales de inteligencia— hace dudar de la práctica común de afirmar la validez concurrente de nuevas pruebas al correlacionarlas con pruebas anteriores. Si las pruebas establecidas sirven como el criterio principal contra el cual se evalúan las nuevas, entonces estas últimas se considerarán válidas sólo en el grado en que se correlacionen con las pruebas antiguas. Dicha práctica conservadora reduce de manera drástica la innovación. La definición operacional de inteligencia anula la posibilidad de que pruebas o conceptos de inteligencia novedosos puedan ser superiores a los existentes.

Por tanto, se debe concluir que las definiciones operacionales de inteligencia dejan mucho que desear. En contraste, una **definición real** es aquella que busca explicar la verdadera naturaleza de lo que se define (Robinson, 1950; Sternberg, 1986). Quizás el modo más común —pero de ninguna manera el único— de producir definiciones reales de inteligencia consista en pedirle a expertos en la materia que la definan.

DEFINICIONES DE EXPERTOS ACERCA DE LA INTELIGENCIA

Investigadores destacados en el área han proporcionado muchas definiciones reales de la inteligencia. A continuación se dan varios ejem-

plos, parafraseados ligeramente para darles consistencia editorial. El lector observará que muchas de estas definiciones aparecieron en un simposio inicial, pero que aún goza de influencia, "*Intelligence and its Measurement*", publicado en el *Journal of Educational Psychology* (Thorndike, 1921). Otras definiciones provienen de una actualización moderna de este primer simposio, "*What is Intelligence?*", editado por Sternberg y Detterman (1986). La inteligencia se ha definido como sigue:

Spearman (1904, 1923): una capacidad general que implica principalmente la deducción de relaciones y correlatos.

Binet y Simon (1905): la capacidad para juzgar bien, para comprender bien y para razonar bien.

Terman (1916): la capacidad para formar conceptos y para comprender su importancia.

Pintner (1921): la capacidad del individuo para adaptarse de manera adecuada a las situaciones relativamente nuevas de la vida.

Thorndike (1921): el poder de dar buenas respuestas, desde el punto de vista de la verdad o el hecho.

Thurstone (1921): la capacidad para inhibir las adaptaciones instintivas, para imaginar de manera flexible diferentes respuestas y para realizar adaptaciones instintivas modificadas en conducta manifiesta.

Wechsler (1939): el conjunto o capacidad global del individuo para actuar de manera propositiva, pensar de manera racional y enfrentarse de manera efectiva con el ambiente.

Humphreys (1971): el repertorio completo de habilidades, conocimiento, inclinación para el aprendizaje y tendencias a la generalización adquiridos, considerados de naturaleza intelectual y que están disponibles en un momento determinado.

Piaget (1972): un término genérico que indica las formas superiores de organización o equilibrio de la estructura cognitiva que se utilizan para la adaptación al ambiente físico y social.

Sternberg (1985a, 1986): la capacidad mental para automatizar el procesamiento de información y para emitir en un contexto la conducta adecuada en respuesta a situaciones novedosas; la inteligencia también incluye metacomponentes, componentes de ejecución y de adquisición del conocimiento (los cuales se analizan después).

Eysenck (1986): transmisión correcta de la información a través de la corteza.

Gardner (1986): la capacidad o habilidad para resolver problemas o para crear productos que se consideran valiosos dentro de uno o más entornos culturales.

Ceci (1994): capacidades innatas múltiples que dan un rango de posibilidades; éstas se desarrollan (o dejan de desarrollarse o se desarrollan y después se atrofian) dependiendo de la motivación y exposición a experiencias educativas relevantes.

La lista de definiciones anterior es representativa, aunque no definitiva ni exhaustiva. En primer lugar, ésta es exclusivamente occidental y omite varios conceptos transculturales de la inteligencia. Por ejemplo, los conceptos orientales acerca de ésta enfatizan la benevolencia, humildad, libertad con respecto a las normas convencionales de juicio y la realización de aquello que es correcto como partes esenciales de la inteligencia. Muchos conceptos africanos de la inteligencia dan un enorme énfasis a sus aspectos sociales, como el mantenimiento de relaciones armoniosas y estables dentro del grupo (Sternberg y Kaufman, 1998). El lector puede consultar a Bracken y Fagan (1990), Sternberg (1994) y Sternberg y Detterman (1986) para ideas adicionales. Seguramente esta muestra de perspectivas es suficiente para demostrar que parece haber tantas definiciones de la inteligencia ¡como expertos dispuestos a definirla!

A pesar de la diversidad de puntos de vista, dos temas se repiten una y otra vez en las definiciones. En términos amplios, los expertos tien-

den a concordar en que la inteligencia es 1) la capacidad para aprender de la experiencia y 2) la capacidad para adaptarse al propio ambiente. El hecho de que tanto el aprendizaje como la adaptación sean importantes para la inteligencia se destaca en gran medida en ciertos casos de discapacidad mental, donde las personas no poseen una u otra capacidad en grado suficiente (ejemplo de caso 5–1).

¿Qué tan bien captan las pruebas de inteligencia la perspectiva de los expertos acerca de que la inteligencia consiste en el aprendizaje a partir de la experiencia y de la adaptación al ambiente? El lector debe mantener en mente esta pregunta a medida que continúe la revisión de las principales pruebas de inteligencia en los temas que siguen. Con seguridad existe causa para preocuparse: muy pocas pruebas de inteligencia contemporáneas parecen requerir que la persona evaluada aprenda algo nuevo o se adapte a una nueva situación como parte esencial del proceso de examen. En el mejor de los casos, las principales pruebas actuales proporcionan medidas indirectas de las capacidades para aprender y adaptarse. Qué tan bien captan estas dimensiones es una pregunta empírica que debe demostrarse a través de investigación de la validez.

CONCEPTOS DE LEGOS Y EXPERTOS ACERCA DE LA INTELIGENCIA

Otro enfoque para comprender un constructo consiste en estudiar su significado popular. Este método es más científico de lo que podría parecer. Las palabras tienen un significado común en el sentido de que ayudan a proporcionar una imagen eficaz de las transacciones cotidianas. Si los legos pueden estar de acuerdo con su significado, un constructo como la inteligencia es, en cierto sentido, algo "real" y, por tanto, potencialmente útil. En tal caso, preguntar a personas en la calle "¿qué significa para usted la palabra inteligencia?", tiene mucho a su favor.

Sternberg, Conway, Ketron y Bernstein (1981) llevaron a cabo una serie de estudios para investigar los conceptos de los adultos estadounidenses acerca de la inteligencia. En el primer estudio, se le pidió a personas que entraban a una estación de trenes, a un supermercado o que estudiaban en una biblioteca universitaria, que listaran las conductas características de diversos tipos de inteligencia. En un segundo estudio —el único que se analiza aquí— tanto legos como expertos (principalmente psicólogos académicos) calificaron la importancia de estas conductas para su concepto de la persona con "inteligencia ideal".

Las conductas centrales para los conceptos de expertos y legos acerca de la inteligencia resultaron ser muy similares, aunque no idénticas. En orden de importancia, los expertos consideraron que la inteligencia verbal, la capacidad de solución de problemas y la inteligencia práctica son importantes para la inteligencia. Los legos consideraron a la capacidad de solución de problemas, la capacidad verbal y la competencia social como factores fundamentales de inteligencia. Por supuesto, las opiniones no fueron unánimes; estos conceptos representan el punto de vista consensual de cada grupo. Los componentes de la inteligencia y sus elementos descriptivos esenciales se muestran en el cuadro 5–1.

En sus conceptos de inteligencia, los expertos enfatizaron la capacidad verbal más que la solución de problemas; mientras los legos invirtieron estas prioridades. No obstante, ambos grupos consideraron que estos dos elementos son aspectos esenciales de la inteligencia. Como verá el lector, la mayor parte de las pruebas de inteligencia también acentúan estas dos capacidades. Ejemplos prototípicos serían vocabulario (capacidad verbal) y diseño con cubos (solución de problemas) de las escalas Wechsler, las cuales se analizan posteriormente. Por tanto, puede verse que los conceptos cotidianos de inteligencia se reflejan, en parte, de manera bastante fiel en las pruebas modernas de inteligencia.

También es evidente cierto desacuerdo entre expertos y legos. Los primeros consideran que la inteligencia práctica (evaluación de situaciones, determinación de cómo lograr metas, conciencia e interés en el mundo) es un componente esencial de la inteligencia.

EJEMPLO DE CASO 5–1
APRENDIZAJE Y ADAPTACIÓN COMO FUNCIONES ELEMENTALES DE LA INTELIGENCIA

Es frecuente que las personas con discapacidad mental demuestren la importancia del aprendizaje por experiencia y de la adaptación al ambiente como ingredientes clave de la inteligencia. Considérese la historia clínica de un vendedor de periódicos de 61 años de edad, con retraso mental moderado, quien era muy conocido para los especialistas en salud mental de la localidad. Era un hombre interesante, por no decir excéntrico, que almacenaba alimentos enlatados en su refrigerador y maldecía a los trabajadores de la beneficencia que pasaban a verlo para verificar su situación. A pesar de su necesidad de apoyo económico de una dependencia del estado, era ferozmente independiente y administraba su propio hogar con mínima supervisión de los encargados de su caso. Así, en muchos sentidos mantenía una adaptación tenue con su ambiente. Para ganar el ingreso adicional que mucho necesitaba, vendía un diario local con un costo de 25 centavos en un quiosco de la calle. Reconocía que la moneda de 25 centavos era el pago adecuado y había aprendido a dar tres monedas de 25 centavos como cambio de un billete de un dólar. Se negaba a aceptar otras formas de pago, un arreglo que sus clientes podían aceptar. Pero un día aumentó el precio del periódico a 35 centavos y el vendedor de periódicos, que sufría retraso mental, se vio forzado a manejar monedas de 10 y 1 centavo, así como las de 25 centavos y billetes de dólar. La cantidad de aprendizaje que se requería por este leve cambio en las demandas del ambiente superó sus capacidades intelectuales y, desgraciadamente, pronto quedó fuera del negocio. Sus intentos fallidos resaltan los ingredientes esenciales de la inteligencia: el aprendizaje a partir de la experiencia y la adaptación al ambiente.

Cuadro 5–1. Factores y elementos de muestra que subyacen a los conceptos de inteligencia de legos y expertos

Legos	Expertos
Capacidad de solución de problemas prácticos	**Inteligencia verbal**
Razona de manera lógica y correcta	Ostenta un buen vocabulario
Identifica las conexicones entre ideas	Lee con elevada comprensión
Puede ver todos los aspectos de un problema	Tiene curiosidad
Conserva la mente abierta	Es inquisitivo en un sentido intelectual
Capacidad verbal	**Capacidad de solución de problemas**
Habla con claridad y de manera articulada	Capaz de aplicar el conocimiento a los proble-
Tiene fluidez verbal	mas que se le presentan
Tiene buena conversación	Toma buenas decisiones
Está bien informado sobre su campo particular	Plantea los problemas de manera óptima
de conocimientos	Tiene sentido común
Competencia social	**Inteligencia práctica**
Acepta a los demás por lo que son	Aprovecha bien las situaciones
Admite sus errores	Determina cómo lograr las metas
Tiene interés en el mundo en general	Tiene conciencia del mundo
Llega a tiempo a sus citas	Muestra interés por el mundo en general

Nota: En cada factor sólo se enumeran los elementos con las cargas más altas. Los investigadores determinaron los nombres de los factores.

Fuente: Reproducido con autorización de Sternberg, R. J., Conway, B. E.; Ketron, J. L. y Bernstein, M. (1981). People's conceptions of intelligence. *Journal of Personality and Social Psychology, 41*, 37-55.

mientras los legos identifican a la competencia social (aceptar a los otros individuos como son, admitir los errores, la puntualidad y el interés en el mundo) como un tercer componente. Sin embargo, estas dos nominaciones comparten una propiedad: las pruebas contemporáneas no intentan medir ya sea la inteligencia práctica o la competencia social. Esto refleja, en parte las dificultades psicométricas que se enfrentan en el diseño de reactivos de una prueba pertinente para estas áreas de contenido. No obstante, la principal razón por la que las pruebas de inteligencia no miden la inteligencia práctica o la competencia social es la inercia: los autores de pruebas han aceptado a ciegas los conceptos históricamente incompletos acerca de la inteligencia. Hasta hace poco tiempo, el desarrollo de estas pruebas había sido una cuestión conservadora, con poco cambio desde los tiempos de Binet y de las pruebas Army Alfa y Beta dirigidas a los reclutas de la primera Guerra Mundial. Sin embargo, existen algunas señales de que es posible que las prácticas de prueba evolucionen en próximas fechas, con el desarrollo de instrumentos innovadores. Por ejemplo, Sternberg y colaboradores han propuesto pruebas innovadoras basadas en su modelo de inteligencia. Otro instrumento interesante, basado en un nuevo modelo de inteligencia, es el *Everyday Problem Solving Inventory* (Inventario de solución de problemas cotidianos; Cornelius y Caspi, 1987). En esta prueba, los examinandos deben indicar su respuesta típica a problemas cotidianos como no llevar dinero, la chequera o una tarjeta de crédito cuando invitan a comer a un amigo.

Se volverá ahora la atención hacia una revisión de las principales teorías de la inteligencia. Como recordatorio, debe decirse que la justificación para revisar las teorías consiste en ilustrar la manera en que éstas han influido en la estructura y contenido de las pruebas de inteligencia. Además, la validez de constructo de las pruebas de CI depende del grado en que encarnen teorías específicas de la inteligencia, de modo que también es pertinente una revisión de las teorías para la validación de pruebas.

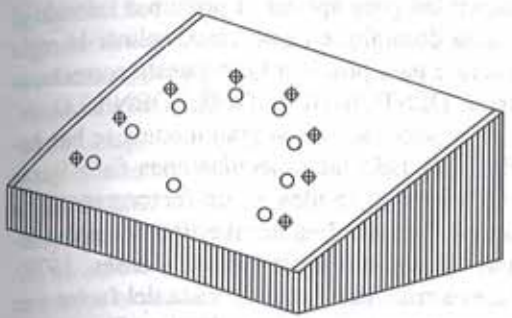
TEORÍAS DE LA INTELIGENCIA

GALTON Y LA AGUDEZA SENSORIAL

Las primeras teorías de la inteligencia surgieron en la era de los instrumentos de bronce de la psicología, a principios del siglo XX. El lector recordará del Tema 1A que Sir Francis Galton y su discípulo, J. McKeen Cattell, consideraban que la inteligencia se sustenta en las capacidades sensoriales agudas. Esta suposición incompleta y engañosa se basó en una premisa plausible:

> La única información que llega a nosotros acerca de los acontecimientos externos parece atravesar por las avenidas de nuestros sentidos; y entre más perceptivos sean los sentidos de la diferencia, mayor es el campo sobre el que puede actuar nuestro juicio e inteligencia (Galton, 1883).

La teoría de la agudeza sensorial acerca de la inteligencia, que promovieron Galton y Cattell, resultó ser en gran medida un callejón sin salida en el sentido psicométrico. Sin embargo, hay vestigios de este enfoque en los análisis cronométricos actuales, como el del aparato de Tiempo de reacción-Tiempo de movimiento (TR-TM), un método experimental que favorece Jensen (1980) para el estudio de la inteligencia sin influencia de la cultura (figura 5–1). En los estudios de TR-TM, se instruye al participante para que coloque el dedo índice de la mano preferida en el botón de inicio; después se hace sonar una señal auditiva a la que le sigue (en 1 a 4 segundos) el encendido de 1 de 8 luces verdes, que el individuo debe apagar con la mayor velocidad posible al tocar el botón del microinterruptor directamente debajo de ella. El TR es el tiempo que requiere el participante para retirar su dedo del botón de inicio después de que se ha encendido una luz verde. El TM es el intervalo entre el que retira su dedo del botón de inicio y toca aquel que apaga la luz verde. Jensen (1980) informó que los índices de TR y de TM se correlacionan en niveles tan altos como .50 con pruebas psicométricas tra-



Nota: Los círculos blancos son los interruptores; los círculos con una cruz son las señales luminosas verdes.

Figura 5–1. Aparato de tiempo de reacción-tiempo de movimiento.
Fuente: Reimpreso con autorización de Jensen, A.R. (1980). *Bias in mental testing.* Nueva York: Free Press. Copyright © 1980 de Arthur R. Jensen. Reimpreso con autorización de Free Press, una división de Simon & Schuster.

dicionales de la inteligencia.* P. A. Vernon también ha informado sobre relaciones sustanciales —tan altas como .70 para correlaciones múltiples— entre medidas de velocidad de procesamiento del tipo de TR y medidas tradicionales de inteligencia (Vernon, 1994; Vernon y Mori, 1990). Estos descubrimientos sugieren que las medidas de velocidad de procesamiento como el TR podrían ser una adición útil para las baterías estandarizadas de pruebas de inteligencia. En general, los autores de pruebas se han resistido a las implicaciones de esta línea de investigación.

SPEARMAN Y EL FACTOR *G*

Con base en un amplio estudio de los patrones de correlaciones entre diversas pruebas de capacidad intelectual y sensorial, Charles Spearman (1904, 1923, 1927) propuso que la inteligencia consistía en dos tipos de factores; un solo **factor general**, *g*, y numerosos **factores específicos**, s_1, s_2, s_3, y así sucesivamente. Como adición importante para su teoría, Spearman ayudó a inventar el análisis factorial como auxiliar para su investigación sobre la naturaleza de la inteligencia. Este autor utilizó esta técnica estadística para distinguir el número de factores subyacentes independientes que deben existir para explicar las correlaciones observadas entre un gran número de pruebas.

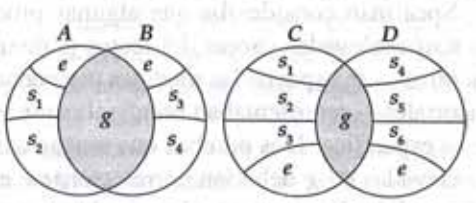
Desde la perspectiva de Spearman, el desempeño de una persona evaluada con cualquier prueba o subprueba homogénea de la capacidad intelectual estaba determinado principalmente por dos influencias: *g*, el factor general dominante, y *s*, un factor específico de esa prueba o subprueba. (Un factor de error, *e*, también puede influir en las puntuaciones, pero Spearman buscó reducir al mínimo esta influencia al utilizar instrumentos sumamente confiables.) Debido a que el factor específico *s* era diferente en cada prueba o subprueba intelectual, y en general tenía menos influencia que *g* en la determinación del nivel de desempeño, Spearman expresó menos interés en estudiarlo. Se concentró principalmente en la definición de la naturaleza de *g*, la cual vinculaba con una "energía" o "fuerza" que surte en común a toda la corteza. En contraste, consideraba que *s*, el factor específico, tenía un sustrato fisiológico localizado en el grupo de neuronas que cubren el tipo particular de operación mental que demanda una prueba o subprueba. Spearman (1923) escribió que: "Así, estos grupos neuronales pueden funcionar como maquinarias opcionales a las que el suministro común de energía puede distribuirse de manera alternativa".

Spearman consideraba que algunas pruebas tenían elevadas cargas del factor *g*, mientras otras —en especial las medidas puramente sensoriales— representaban principalmente un factor específico. Dos pruebas que tengan cargas elevadas de *g* deberían correlacionarse en gran medida. En contraste, las pruebas psicológicas no saturadas con *g* deberían mostrar una correlación mínima entre sí. Gran parte de la investigación de Spearman se dirigió a demostrar la veracidad de estas proposiciones básí-

* De hecho, el coeficiente bruto de correlación es negativo porque los tiempo de reacción **más rápidos** (puntuaciones numéricas **menores**) se relacionan con puntuaciones más altas de inteligencia.

cas derivadas de su teoría. En la figura 5–2 se ilustran de manera gráfica estos conceptos; cada círculo representa una prueba de inteligencia y el grado de superposición entre los círculos indica la fortaleza de la correlación. Obsérvese que **A** y **B**, cada una con cargas elevadas de *g*, se correlacionan en gran medida. Las pruebas **C** y **D** tienen cargas débiles en *g* y, en consecuencia, no se correlacionan bien.

Spearman (1923) creía que las diferencias individuales en *g* se reflejaban de manera más directa en la capacidad para utilizar tres principios de la cognición: aprehensión de la experiencia, educción de relaciones y educción de correlaciones. Incidentalmente, el término **educción**, que se utiliza muy poco, se refiere al proceso de comprender las cosas. Estos tres principios pueden explicarse al examinar la manera en que se resuelven analogías del tipo A:B::C:?; es decir, A es a B, como C es a? Un ejemplo sencillo podría ser MARTILLO: CLAVO::DESTORNILLADOR:? Para resolver esta analogía, primero debe percibirse y comprenderse cada término con base en la experiencia pasada; es decir, debe tenerse aprehensión de la experiencia. Si no se tiene idea de lo que son un martillo, un clavo y un destornillador, existe muy poca posibilidad de que pueda completarse la analogía de manera correcta. A continuación, debe inferirse la relación entre los primeros dos términos de la analogía, en este caso MARTILLO y CLAVO. Mediante una frase un tanto ampulosa, Spearman se refirió a la relación entre dos conceptos como educción de relaciones. El paso final, la educción de correlatos, se refiere a la capacidad para aplicar el principio inferido al nuevo dominio, en este caso, aplicar la regla inferida para producir la respuesta correcta, es decir, DESTORNILLADOR:TORNILLO.



Nota: Las pruebas A y B tienen una fuerte correlación, mientras C y D tienen una correlación débil. Véase texto.

Figura 5–2. Teoría de Spearman sobre los dos factores de la inteligencia.

A pesar de que en gran medida se han hecho a un lado las especulaciones fisiológicas de Spearman, la idea de un factor general ha sido tema central en la investigación sobre inteligencia y aún sigue vigente (Jensen, 1979). La exactitud del punto de vista del factor *g* es más que una cuestión académica. Si es cierto que un factor único, dominante y general es la fuente esencial de la inteligencia, entonces los esfuerzos psicométricos por producir subpruebas factorialmente puras (p. ej., medir la comprensión verbal, organización perceptual, memoria a corto plazo, etc.) están muy equivocados. En la medida en que Spearman esté en lo correcto, los autores de pruebas deberían evitar la derivación de subpruebas y concentrarse en producir una prueba que capture de mejor manera el factor general.

La cuestión más difícil que enfrenta la teoría de los dos factores de Spearman es la existencia de los factores grupales. Ya desde 1906, este autor y sus contemporáneos observaron que pruebas relativamente diferentes podían tener correlaciones más altas que los valores predichos a partir de sus respectivas cargas de *g* (Brody y Brody, 1976). Sus descubrimientos hicieron surgir la posibilidad de que un grupo de medidas diversas pudiera compartir una capacidad unitaria diferente a *g*. Por ejemplo, varias pruebas podrían compartir un factor unitario común de memorización que se encontrara a medio camino entre el factor *g* y los diversos factores *s* únicos de cada prueba. Por supuesto, la existencia de factores grupales es incompatible con la meticulosa teoría de Spearman acerca de dos factores.

THURSTONE Y LAS HABILIDADES MENTALES PRIMARIAS

Thurstone (1931) desarrolló procedimientos de análisis factorial capaces de buscar matrices

de correlación para la existencia de factores grupales. Sus métodos permitían que un investigador descubriera a nivel empírico el número de factores presentes en una matriz y que definiera cada factor en términos de las pruebas que tenían cargas en él. En su análisis de la manera en que se correlacionaban entre sí diferentes tipos de pruebas intelectuales, Thurstone concluyó que varios factores grupales amplios —y no un solo factor general— podían representar la mejor explicación de los resultados empíricos. En diversos momentos de su carrera de investigación, Thurstone propuso aproximadamente una docena de factores diferentes. Sólo siete de éstos se han corroborado con frecuencia (Thurstone, 1938; Thurstone y Thurstone, 1941) y se les ha denominado **habilidades mentales primarias** (HMP); son las siguientes:

- Comprensión verbal: la mejor medida es el vocabulario, pero esta habilidad también participa en la comprensión de lectura y en analogías verbales.
- Fluidez de palabra: se mide con pruebas como anagramas o enunciación rápida de palabras dentro de una categoría dada (p. ej., palabras que comienzan con la letra s).
- Número: sinónimo de velocidad y precisión en cálculos aritméticos sencillos.
- Espacio: capacidad para visualizar la manera en que se vería un objeto tridimensional si se le girara o desarmara.
- Memoria asociativa: habilidad en tareas de memoria de repetición, como el aprendizaje de pares asociados de elementos sin relación.
- Velocidad perceptual: participa en tareas sencillas de carácter menor, como búsqueda de semejanzas y diferencias en detalles visuales.
- Razonamiento inductivo: las mejores medidas de este factor implican el hallazgo de una regla, como en una prueba de terminación de una serie de números.

Thurstone (1938) publicó la Prueba de Habilidades Mentales Primarias, que constaba de varias subpruebas independientes diseñadas cada una para medir una HMP. No obstante, posteriormente reconoció que sus habilidades mentales primarias se correlacionaban de manera moderada entre sí, lo cual probaba la existencia de uno o más factores de segundo orden. Por último, Thurstone reconoció la existencia de *g* como factor de primer orden. Para ese momento, Spearman había admitido la existencia de factores grupales que representaban capacidades especiales, y se hizo evidente que las diferencias entre Spearman y Thurstone eran en mayor medida una cuestión de énfasis (Brody y Brody, 1976). Spearman continuó creyendo que *g* era el principal determinante de las correlaciones entre puntuaciones de prueba y asignó una función menor a los factores grupales. Thurstone invirtió estas prioridades.

P. E. Vernon (1950) proporcionó una reconciliación entre estas dos perspectivas al proponer una teoría de factores grupales jerárquicos. Desde su punto de vista, *g* era un factor singular al tope de una jerarquía que incluía dos factores grupales mayores, llamados verbal-educativo (V:ed) y práctico-mecánico-espacial-físico (k:m). Debajo de estos dos factores grupales principales se encontraban varios factores grupales menores semejantes a las HMP de Thurstone; los factores específicos ocupaban la parte inferior de la jerarquía (figura 5–3).

El análisis de Thurstone sobre las HMP aún influye en el desarrollo de pruebas. Schaie (1983, 1985) ha revisado y modificado la Prueba de Habilidades Mentales Primarias y utilizó estas medidas en un estudio longitudinal de enorme influencia acerca de la inteligencia en

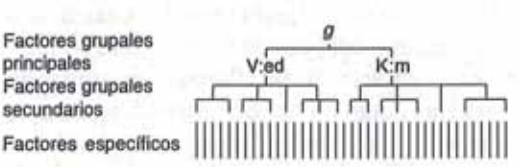


Figura 5–3. Teoría de la inteligencia de Vernon sobre los factores grupales jerárquicos.
Fuente: Reimpreso con autorización de Vernon, P. E. (1950) *The structure of human abilities*. Londres: Methuen.

el adulto. Si la inteligencia fuera principalmente una cuestión de *g*, entonces los factores grupales deberían cambiar aproximadamente en la misma proporción a medida que se envejece. En apoyo al enfoque de los factores grupales para el examen intelectual, Schaie (1983) informa que algunas HMP muestran poca disminución relacionada con la edad (Comprensión verbal, Fluidez de palabra, Razonamiento inductivo), mientras otras disminuyen de manera más rápida en la vejez (Espacio, Número). Así, pueden existir razones prácticas y realistas para informar de factores grupales y no condensar toda la inteligencia en un solo factor general.

R. CATTELL Y LA DISTINCIÓN ENTRE INTELIGENCIA FLUIDA/ CRISTALIZADA

Raymond Cattell (1941, 1971) propuso una célebre teoría acerca de la estructura de la inteligencia que John Horn (1968, 1994) revisó y amplió. Como ocurrió con sus antecesores, Cattell y Horn utilizaron el análisis factorial para estudiar la estructura de la inteligencia. No obstante, en vez de encontrar un solo factor general o una media docena de factores grupales, identificaron dos factores principales que denominaron inteligencia fluida (g_f) e inteligencia cristalizada (g_c).

La **inteligencia fluida** es primordialmente un tipo de eficiencia mental no verbal y relativamente independiente de la cultura. Se relaciona con la capacidad inherente de la persona para aprender y resolver problemas. En ese sentido, la inteligencia fluida se utiliza cuando una tarea requiere adaptación a una situación nueva. En contraste, la **inteligencia cristalizada** representa aquello que se ha aprendido a través de la aplicación de la inteligencia fluida en entornos culturales (p. ej., el aprendizaje de álgebra en la escuela). La inteligencia cristalizada depende principalmente de la cultura y se le utiliza para tareas que requieren una respuesta aprendida o habitual. Dado que ésta surge cuando la inteligencia fluida se aplica a los productos culturales, se esperaría que estos dos tipos de inteligencia estuvieran correlacionados. De hecho, es común encontrar que las medidas de inteligencia cristalizada y fluida se correlacionen de manera moderada ($r = .5$).

Las habilidades que componen la inteligencia fluida son no verbales y no dependen mucho de la exposición a una cultura específica. Por estas razones, Cattell (1940) consideraba que las medidas de este tipo de inteligencia eran independientes de la cultura. Con base en esta suposición, diseñó la *Culture Fair Intelligence Test*, CFIT (Prueba de Inteligencia Culturalmente Justa) como un intento por eliminar el prejuicio cultural en las pruebas. Por supuesto, llamar culturalmente justa a una prueba no necesariamente hace que lo sea. De hecho, el objetivo de lograr una prueba de inteligencia que sea independiente por completo de la cultura ha resultado ser esquiva. En el Tema 6B, Pruebas grupales de inteligencia, se analiza con mayor detalle el tema de la CFIT.

En versiones posteriores de la teoría de la inteligencia fluida/cristalizada, Cattell (1971) y Horn (1982, 1994) aumentaron y elaboraron los conceptos que se analizaron antes. En la actualidad, sería mejor considerar su enfoque como una teoría de inteligencias múltiples, pero la división de g_f-g_c se ha vuelto tan conocida que no será fácil eliminarla. En sus últimas revisiones, los autores han propuesto un modelo jerárquico, entremezclado, de la inteligencia, con los componentes fluido y cristalizado en la parte superior. Estas habilidades tienen subordinados diversos componentes identificados de la inteligencia, que incluyen organización visual, velocidad perceptual, organización auditiva, varias capacidades de memoria, así como componentes específicos de recepción sensorial. El modelo revisado es difícil y los lectores interesados deberían consultar a Horn (1994).

PIAGET Y LA ADAPTACIÓN

El psicólogo suizo Jean Piaget (1896-1980) elaboró una teoría del desarrollo cognitivo que tiene varias implicaciones para el diseño de

pruebas de inteligencia infantil (Ginsburg y Opper, 1988). Piaget (1926, 1952, 1972) utilizó entrevistas y pruebas informales con niños para desarrollar una serie de puntos de vista provocativos y revolucionarios acerca del desarrollo intelectual. Su nueva perspectiva incluía los siguientes elementos:

- El pensamiento infantil es cualitativamente diferente del pensamiento adulto.
- Las estructuras psicológicas denominadas **esquemas** son la base principal para la obtención de conocimiento nuevo acerca del mundo.
- Se pueden identificar cuatro etapas del desarrollo cognitivo.

A continuación se examinan con mayor detalle cada uno de estos puntos.

A través del estudio del desarrollo de la conservación, Piaget concluyó que el niño construye su concepto del mundo de manera fundamentalmente diferente a la perspectiva adulta. La **conservación** se refiere a la conciencia de que las cantidades físicas no cambian cuando se les altera de modo superficial en su apariencia. Por ejemplo, la mayoría de los adultos saben que dos filas iguales de 10 monedas de un centavo siguen siendo equivalentes aunque una de las filas se disperse —poseen conservación de número. Pero será fácil que un niño pequeño se engañe por el cambio aparente, y es posible que insista en que la segunda fila tiene ahora más monedas. De manera similar, se puede mostrar que los niños pequeños no poseen conservación de una cantidad, materia, peso o volumen continuos.

A fin de explicar cómo obtienen los lactantes y niños el conocimiento nuevo acerca del mundo, Piaget sugirió que forman estructuras psicológicas denominadas **esquemas**. Un esquema es un patrón organizado de conducta o una estructura mental bien definida que conduce al conocimiento de cómo se hacen las cosas. Quizás unos cuantos ejemplos ayuden a aclarar este difícil concepto. Los lactantes poseen esquemas que tienen una naturaleza principalmente sensoriomotora, como asir y tirar, que permite a un bebé tomar el objeto deseado y llevárselo a la boca. A medida que crece, se añaden estructuras mentales al conjunto de esquemas sensoriomotores. Por ejemplo, los adolescentes poseen, en términos generales, el esquema de alfabetización que les permite encontrar una palabra en un diccionario mediante la aplicación repetida de la sencilla regla de que las palabras se localizan en orden alfabético por la primera letra, después la segunda y así sucesivamente.

El genio de Piaget se encuentra en su sugerencia del mecanismo a través del cual evolucionan los esquemas hacia niveles de complejidad cada vez mayores, que se trasforman de esa manera en el nivel más maduro de habilidad intelectual que se observa en la mayoría de los adultos. Este mecanismo se conoce como proceso de **equilibración**. Para comprender este proceso, el lector necesita conocer tres conceptos piagetianos adicionales: asimilación, acomodación y equilibrio.

La **asimilación** es la aplicación de un esquema a un objeto, persona o acontecimiento. Por ejemplo, la asimilación participa cuando un lactante emplea el esquema de asir y tirar para tomar una sonaja y llevársela a la boca. Si la asimilación funciona para lograr las metas deseadas por la persona, se presenta un estado de armonía o equilibrio, pero, ¿qué sucede si la aplicación de un esquema no funciona? Supóngase que el esquema de asir y tirar no tiene éxito porque la sonaja del niño se engancha en las barras verticales de la cuna en el momento en que éste trata de llevarse el juguete a la boca. Surgirá entonces un estado de tensión dinámica que requiere que el niño ajuste su esquema para que funcione. La adaptación de un esquema fallido para que funcione se denomina **acomodación**. En el ejemplo del lactante, el esquema podría modificarse y volverse un esquema de asir, tirar y girar. Si el esquema modificado funciona y le permite al lactante llevarse la sonaja a la boca, existe de nuevo un estado de equilibrio. Véase la distinción entre equilibrio, estado de armonía temporal y equilibración, el proceso completo de asimilación, acomodación y equilibrio. Piaget consideraba

que la búsqueda del equilibrio era una característica inherente del ser humano.

Este autor también propuso cuatro etapas del desarrollo cognitivo. Según su perspectiva, cada etapa es cualitativamente diferente de las otras y se caracteriza por patrones característicos de pensamiento (cuadro 5–2). En el tema 5B, Evaluación de las capacidades de lactantes y preescolares, se analiza una prueba para lactantes basada en un análisis piagetiano del desarrollo cognitivo. En general, las pruebas basadas en estos conceptos buscan valorar si un niño ha superado ciertos hitos cognitivos (p. ej., conservación de volumen) propuestos por Piaget.

GUILFORD Y EL MODELO DE LA ESTRUCTURA INTELECTUAL

Después de la segunda Guerra Mundial, J. P. Guilford (1967, 1985) continuó con la búsqueda de los factores de la inteligencia que había iniciado Thurstone. En poco tiempo, Guilford concluyó que el número de habilidades mentales discernibles era mucho mayor que las siete propuestas por Thurstone. En primer lugar, Thurstone había ignorado por completo la categoría de pensamiento creativo, un descuido injustificable desde el punto de vista de Guilford. También encontró que si incluía tipos innovadores de pruebas dentro de las grandes baterías que aplicaba a los individuos, entonces el patrón de correlaciones entre estas pruebas indicaba la existencia de, literalmente, docenas de nuevos factores del intelecto. Lo que es más, Guilford se dio cuenta de que algunos de estos nuevos factores tenían semejanzas recurrentes con respecto a los tipos de procesos mentales implicados, los tipos de información mostrados o la forma que asumían los reactivos de información. Como resultado de estas semejanzas recurrentes en los factores recién descubiertos del intelecto, llegó a la conclusión de que estos factores multitudinarios podían agruparse en un pequeño número de dimensiones principales. Guilford (1967) propuso un elegante modelo de la estructura intelectual (EI) para resumir sus resultados. Concebido en términos visuales, el modelo de EI de Guilford clasifica a las habilidades intelectuales en tres dimensiones denominadas operaciones, contenidos y productos.

Cuadro 5–2. Etapas del desarrollo cognitivo según Piaget

Etapa y rango de edades	Características del pensamiento
Sensoriomotora: nacimiento a los 2 años	Los lactantes experimentan al mundo principalmente a través de sus sentidos y capacidades motoras, actúan como si un objeto dejara de existir cuando no está ante su vista, pero desarrollan permanencia de éste al final de este estadio.
Preoperacional: 2 a 6 años	No se desarrollan aún los conceptos de conservación, pero estos niños comprenden la idea de una relación funcional —por ejemplo, se tira de un cordel para abrir una cortina y entre más se tira de él, más se abre la cortina. También se desarrolla la capacidad para simbolizar mentalmente las cosas con palabras e imágenes.
Operaciones concretas: 7 a 12 años	Por lo común, los niños desarrollan conservación y demuestran capacidades limitadas de razonamiento lógico. Por ejemplo, se desarrolla el concepto de reversibilidad —el conocimiento de que una acción puede revertir o negar a otra.
Operaciones formales: 12 años en adelante	En general en esta etapa se desarrolla la solución sistemática de los problemas que se asocian con el pensamiento adulto. Existe mayor capacidad para generar hipótesis y someterlas a prueba.

Por **operaciones**, Guilford tenía en mente el tipo de operación intelectual que requiere la prueba. La mayor parte de los reactivos enfatiza sólo una de las operaciones enumeradas aquí:

Cognición	Descubrir, entender o comprender.
Memoria	Introducción de los reactivos de información a la memoria, como en el caso de una serie de números.
Producción divergente	Recuperar de la memoria los reactivos de una clase específica, como nombrar objetos comestibles o no.
Producción convergente	Recuperación de la memoria de un reactivo correcto, como en el caso de una palabra para un crucigrama.
Evaluación	Determinar qué tan bien satisface un cierto reactivo de información requisitos lógicos específicos.

Contenidos se refiere a la naturaleza de los materiales o información presentados a la persona evaluada. Las cinco categorías de contenido son las siguientes:

Visual	Imágenes presentadas a los ojos.
Auditivo	Sonidos presentados a los oídos.
Simbólico	Como en el caso de símbolos matemáticos que representan algo.
Semántico	Significados, en general símbolos, de palabras.
Conductual	La capacidad para comprender el estado mental y conducta de otras personas.

La tercera dimensión en el modelo de Guilford, **productos,** se refiere a los diferentes tipos de estructuras mentales que debe producir el cerebro para obtener una respuesta correcta. Los seis tipos de productos son los siguientes:

Unidad	Una entidad singular que tiene una combinación única de propiedades o atributos.
Clase	Aquello que tienen en común las unidades similares, como en un conjunto de triángulos o de sonidos con tonos altos.
Relación	Una conexión observada entre dos reactivos, como dos tonos con una separación de una octava.
Sistema	Tres o más reactivos que forman un todo reconocible, como en una melodía o en un plan para una secuencia de acciones.
Transformación	Un cambio en un reactivo de información, como en el caso de una corrección para un error de ortografía.
Implicación	Lo que implica un reactivo individual, como la expectativa de un trueno después de un rayo.

Así, Guilford (1985) identificó cinco tipos de operaciones, cinco tipos de contenido y seis tipos de productos, para un total de 5 x 5 x 6 o 150 factores del intelecto. Cada combinación de una operación (p. ej., memoria), un contenido (p. ej., simbólico) y un producto (p. ej., unidades) representa un factor diferente del intelecto. Guilford afirma haber verificado más de 100 de estos factores en su investigación.

Con frecuencia se alaba al modelo de la EI por considerar que capta las complejidades de la inteligencia. No obstante, éste también es un talón de Aquiles potencial para la teoría. Considérese un factor del intelecto, la memoria de unidades simbólicas. Una prueba que requiere que el examinando recuerde una serie de dígitos **hablados** (p. ej., Retención de dígitos del WAIS-III) podría captar bastante bien este factor del intelecto. Pero de la misma manera podría lograrlo una prueba **visual** de retención de dígitos y quizá también una prueba análoga con la presentación **táctil** de símbolos, como en el caso de rodillos vibratorios aplicados a la piel. Quizá sea necesario tener un cubo separado para audición, visión y tacto; de tal manera que un modelo ampliado incorporaría 450 factores del intelecto, lo cual con toda seguridad es un número pesado y difícil de manejar.

Aunque parece dudoso que la inteligencia pudiera implicar un número tan grande de capacidades únicas, de todas maneras la perspectiva atomista de Guilford ha provocado que los

autores de pruebas reconsideren y amplíen su comprensión acerca de la inteligencia. Antes de las contribuciones de Guilford, la mayor parte de las pruebas de inteligencia requerían principalmente de la producción convergente —la construcción de una sola respuesta correcta para una situación estímulo. Guilford hizo surgir la fascinante posibilidad de que la **producción divergente**— la creación de numerosas respuestas apropiadas para una sola situación estímulo —también sea un elemento esencial de la conducta inteligente. Así, una indicación como "enumere tantas consecuencias como sea posible de que las nubes tuvieran cuerdas colgando de ellas" (producción divergente), podría evaluar un aspecto de la inteligencia que no se mide a través de las pruebas tradicionales.

TEORÍA DEL PROCESAMIENTO SIMULTÁNEO Y SUCESIVO

Algunos de los conceptos actuales de la inteligencia están en deuda con las investigaciones neuropsicológicas del psicólogo ruso Aleksandr Luria (1902-1977). Luria (1966) se basó primordialmente en estudios sobre casos individuales y en observaciones clínicas de soldados con lesiones cerebrales para llegar a una teoría general del procesamiento cognitivo. El núcleo de su teoría es el siguiente:

> Los análisis muestran que existe fuerte evidencia para distinguir dos tipos básicos de actividad integradora de la corteza cerebral mediante las cuales pueden reflejarse diferentes aspectos del mundo exterior... La primera es la integración de los estímulos individuales que llegan al cerebro en grupos simultáneos y principalmente espaciales, y la segunda es la integración de los estímulos individuales que llegan de manera consecutiva al cerebro en series sucesivas, organizadas en sentido temporal (Luria, 1966).

Dado que este enfoque se centra en la mecánica a través de la cual se procesa la información, con frecuencia se le conoce como teoría del procesamiento de información.

El **procesamiento simultáneo** de la información se caracteriza por la ejecución simultánea de diferentes operaciones mentales. Los tipos de pensamiento y percepción que requieren del análisis espacial, como el dibujo de un cubo, necesitan del procesamiento simultáneo de información. En el dibujo, la persona evaluada debe percibir de manera simultánea la forma general y guiar su mano y dedos en la ejecución de la forma. Un enfoque secuencial para el dibujo de un cubo (si al menos eso fuera posible) sería horriblemente complejo. En efecto, el examinando tendría que dibujar líneas individuales de longitudes y orientaciones angulares sumamente específicas y esperar que todo se alineara. En ausencia de una *gestalt* mental simultánea que guíe el dibujo, casi es seguro que se obtenga una producción distorsionada. Luria descubrió que el procesamiento simultáneo se relaciona con los lóbulos occipital y parietal en la parte trasera del cerebro.

El **procesamiento sucesivo** de la información es necesario para actividades mentales en las que debe seguirse una secuencia apropiada de operaciones. Se encuentra en agudo contraste con el procesamiento simultáneo (como en el dibujo), para el que la secuencia no tiene importancia. El procesamiento sucesivo es necesario para recordar una serie de dígitos, repetir una cadena de palabras (p. ej., zapato, pelota, huevo) y para imitar una serie de movimientos con la mano (puño, palma, puño, puño, palma). Luria localizó el procesamiento sucesivo en el lóbulo temporal y en las regiones frontales cercanas a éste.

La mayor parte de las formas de procesamiento de información requieren de interacción de mecanismos simultáneos y sucesivos. Das (1994) cita el ejemplo de la lectura de una palabra desconocida como **taciturno:**

> Deben reconocerse las letras individuales y eso implica la codificación simultánea. El lector equipara la forma visual de la letra con un diccionario mental y obtiene el nombre para ella. Entonces, han de formarse las secuencias de letras (codificación sucesiva) y combinarse como una sílaba (simultánea). Después la cadena de sílabas debe convertirse en una palabra (sucesiva), reconocerse la palabra (simultánea) y luego armarse un programa de pronunciación (sucesiva), lo cual conduce a la lectura oral (sucesiva y simultánea).

Das admite que esto puede representar una perspectiva simple de lo que ocurre cuando un lector se enfrenta con una palabra. El punto esencial es que el procesamiento de información de nivel superior depende de la interacción de formas específicas, anatómicamente localizables, de procesamiento de información.

El reto de un enfoque simultáneo-sucesivo para la evaluación de la inteligencia consiste en diseñar tareas que detecten formas relativamente puras de cada enfoque de procesamiento de información. Las pruebas que emplean esta estrategia son la Batería Kaufman de Evaluación para Niños (K-ABC), que se analiza en el siguiente tema, y el *Das-Naglieri Cognitive Assmessment System* (Sistema de Evaluación Cognitiva Das-Naglieri; Das y Naglieri, 1993). La batería Das-Naglieri incluye tareas sucesivas que implican enunciación rápida (como "Diga **pan, sol, col** tan rápido como le sea posible en 10 ocasiones") y medidas simultáneas tanto de tareas verbales como no verbales; también evalúa la planificación y atención, lo cual conduce a las siglas PASS (planificación, atención, simultáneo, sucesivo; Das, Naglieri y Kirby, 1994).

INTELIGENCIA COMO CONSTRUCTO BIOLÓGICO

La mayoría de los investigadores han estudiado la inteligencia de manera tradicional al desarrollar pruebas del intelecto y correlacionar las puntuaciones con criterios externos (p. ej., calificaciones escolares) u otros resultados de una prueba. Pero algunos han buscado comprender la naturaleza de la inteligencia al examinar las propiedades del cerebro en sí. Por ejemplo, Hynd y Willis (1985) proporcionaron un excelente estudio de los fundamentos neurológicos de la inteligencia.

Una propiedad importante del cerebro, la cual se requiere para la conducta inteligente, es la actividad eléctrica sincronizada y con un patrón adecuado de las células cerebrales. Las neuronas deben trasmitir impulsos electroquímicos calibrados de manera precisa a fin de que ocurran la sensación, percepción y los procesos superiores de pensamiento. La actividad eléctrica colectiva de las neuronas puede medirse colocando electrodos en el cuero cabelludo de la persona. El registro continuo de la actividad eléctrica muestra fluctuaciones espontáneas a lo largo del tiempo, pero también demuestra patrones predecibles en respuesta a ciertos estímulos. Por ejemplo, se puede medir un potencial evocado al observar el patrón de ondas cerebrales que ocurren en un cuarto de segundo más o menos después de que se enciende una luz ante los ojos de un individuo. En general, el potencial evocado promedio (PEP) se obtiene de cientos de dichos ensayos con una sola persona. De esta manera puede obtenerse un patrón sumamente consistente y distintivo para cualquier individuo.

Ertl y Schafer (1969) se encuentran entre los primeros investigadores que estudiaron los correlatos en ondas cerebrales de la inteligencia. Estos autores descubrieron que la forma de la onda en el PEP tiene más picos y valles en el caso de individuos con CI elevado que en el de aquellos con CI bajo. Eysenck (1982) publicó descubrimientos similares que se reproducen en la figura 5–4. Dos colaboradores de Eysenck, A. E. Hendrickson (1982) y D. E. Hendrickson (1982), señalaron que la longitud total de la forma sinuosa de la onda en el PEP podía utilizarse como índice biológico de la inteligencia. Colocaron un trozo de cordel sobre cada una de las formas de onda del PEP informadas por Ertl y Schafer (1969). Cortaron el inicio y fin de los cordeles, los extendieron de manera firme para formar líneas rectas y después midieron su longitud. Entonces, los investigadores pudieron calcular la correlación entre las longitudes de los cordeles y las puntuaciones publicadas de CI. El resultado fue un impresionante valor de $r = .77$. Esta correlación es tan elevada como aquéllas informadas entre cualesquiera dos pruebas psicométricas de inteligencia. Una medida puramente biológica de la función cerebral (ondas del PEP) resulta ser un excelente mecanismo de predicción de la inteligencia como se mide a través de pruebas tradicionales de CI.

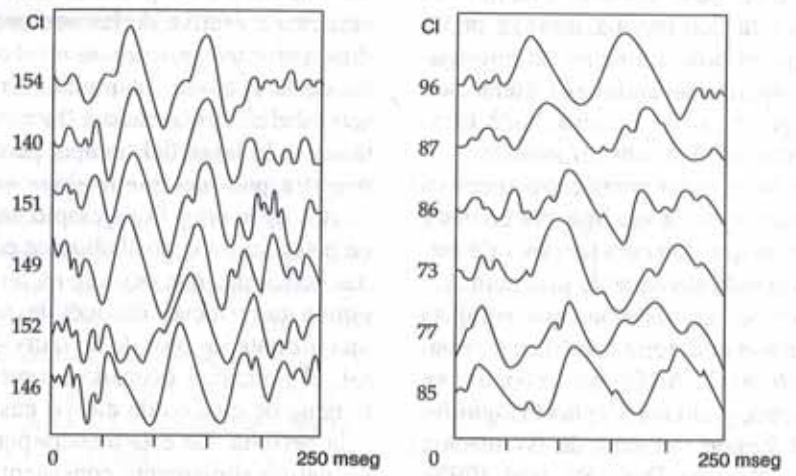


Figura 5–4. Formas de onda del potencial evocado promedio (PEP) para personas con CI alto y bajo.
Fuente: Reproducido con autorización de Eysenck, H. J. (1982). *A model for intelligence.* Heidelberg: Springer. Utilizado con autorización de Springer Publishing Company Inc., 536 Broadway, Nueva York, NY 10012.

A pesar de estos resultados prometedores de la investigación, varios estudiosos continúan mostrándose escépticos acerca de los correlatos electrocorticales de la inteligencia. Las correlaciones surgen sólo bajo ciertas condiciones y los intentos por replicar los resultados no siempre tienen éxito (Eysenck, 1994; Vernon y Mori, 1990). Gale y Edwards (1983) insisten en que los estudios de correlación no son suficientes; es necesaria una orientación más ligada a la teoría que vincule la inteligencia como rasgo con el procesamiento de información a nivel neuronal. Se han hecho intentos por formular una teoría de este tipo (Deary, Hendrickson y Burns, 1987). Éstos y otros estudios similares (p. ej., Shucard y Horn, 1972) sirven para recordar que la inteligencia se encuentra ligada de alguna manera a las propiedades fisiológicas del cerebro, aunque no se comprenda aún la característica biológica precisa que la explica.

Haier y sus colaboradores han seguido una vía diferente en su estudio acerca de la inteligencia biológica (Haier, Nuechterlein, Hazlett y col., 1988; Haier, Siegel, Tang y col., 1992). Estos investigadores midieron las tasas metabólicas de la glucosa cortical, como se revela a través de los análisis con tomografía por emisión de positrones (TEP), de voluntarios que resuelven problemas intelectuales. Las neuronas utilizan glucosa y oxígeno como combustible, de modo que una TEP revelará "puntos calientes" en la mayor parte de las zonas cerebrales activas (donde la glucosa se metaboliza). Un resultado fascinante es que las personas más inteligentes mostraron una **menor** actividad cerebral al momento de resolver problemas de analogías geométricas y cuando jugaban el juego de computadora Tetris, que aquéllas menos inteligentes. Lo que aún es poco claro en esta línea de investigación es la dirección causal: ¿las personas son inteligentes porque utilizan menos glucosa o la usan menos porque son inteligentes? Otra posibilidad es que tanto el CI elevado como el bajo metabolismo de la glucosa se encuentren relacionados con una tercera variable causal (Sternberg y Kaufman, 1998).

GARDNER Y LA TEORÍA DE LAS INTELIGENCIAS MÚLTIPLES

Howard Gardner (1983, 1993) ha propuesto una teoría de las inteligencias múltiples basado de manera aproximada en el estudio de las relaciones cerebro-conducta. Afirma que existen varias inteligencias humanas relativamente independientes, aunque admite que aún no se han establecido en definitiva la naturaleza, grado y número exactos de las inteligencias. Gardner (1983) esboza de la siguiente manera los criterios para una inteligencia autónoma:

- Aislamiento potencial por daño cerebral —la lesión cerebral puede haber destruido, o protegido por aislamiento, la facultad.
- La existencia de individuos excepcionales, como los autistas sabios —la facultad queda particularmente protegida en medio de la mediocridad intelectual.
- Operaciones centrales identificables —la facultad depende de una o más operaciones básicas de procesamiento de información.
- Historia distintiva del desarrollo —la facultad posee una historia de desarrollo identificable que quizá incluya periodos críticos e hitos.
- Admisibilidad evolutiva —aunque esto se encuentra a nivel especulativo, una facultad debe tener antecedentes evolutivos que comparta con otros organismos (p. ej., la organización social de los primates).
- Sustento de la psicología experimental —la facultad surge en estudios de laboratorio de psicología cognitiva.
- Sustento de descubrimientos psicométricos —la facultad se revela en estudios de medición y es susceptible a la medición psicométrica.
- Susceptibilidad para la codificación simbólica —la facultad puede comunicarse a través de símbolos, incluyendo (aunque no limitado a) lenguaje, representación pictórica y matemáticas.

Con base en estos criterios, Gardner (1983, 1993) propone que se han confirmado de manera sustancial las siguientes siete inteligencias naturales: lingüística, lógico-matemática, espacial, musical, corporal-cinestésica, interpersonal e intrapersonal. Tres de estos siete tipos de inteligencia son muy conocidos —lingüística (es decir, verbal), lógico-matemática y espacial— y se han diseñado numerosas pruebas formales para medirlas, de modo que aquí no se discutirá más acerca de ellas. Las otras cuatro variaciones de inteligencia son en cierto modo novedosas y, por tanto, requieren de una presentación más detallada.

La inteligencia corporal-cinestésica incluye los tipos de habilidades que utilizan los atletas, bailarines, mimos, mecanógrafos o cazadores "primitivos". Aunque por lo general las culturas occidentales se muestran reacias a considerar al cuerpo como un tipo de inteligencia, este no es el caso en gran parte del resto del mundo, como tampoco ocurre así en la historia evolutiva. De hecho, las personas que tenían habilidad para evitar a los depredadores, subirse a los árboles, cazar animales y preparar herramientas estaban en mayor probabilidad de sobrevivir y pasar sus genes a generaciones posteriores.

Las inteligencias personales incluyen la capacidad para tener acceso a la propia vida emocional (intrapersonal), así como la de advertir y distinguir los estados de ánimo, temperamentos, motivaciones e intenciones de los demás (interpersonal). Así, la inteligencia personal engloba tanto una versión intrapersonal como una interpersonal. La primera se encuentra en los grandes novelistas que pueden escribir de manera introspectiva acerca de sus sentimientos, mientras la última puede observarse con frecuencia en líderes religiosos y políticos (p. ej., Mahatma Gandhi o Lyndon Johnson) que pueden desentrañar las intenciones y deseos de otras personas y utilizar esa información para influir sobre ellas y formar alianzas útiles.

La inteligencia musical es quizá la menos comprendida de las inteligencias de Gardner. Las personas con buena inteligencia musical aprenden con facilidad a tocar un instrumento o a escribir sus propias composiciones. Aunque el conocimiento de los aspectos estructurales de la melodía, ritmo y timbre es importante para la inteligencia musical, Gardner señala que

muchos expertos consideran que los aspectos afectivos o emocionales de la música son el elemento central. Este autor considera que cuando finalmente se descifren los fundamentos neurológicos de la música, se tendrá "una explicación de la manera en que los factores emocionales y motivacionales se encuentran entrelazados con aquéllos puramente perceptuales" (Gardner, 1983).

El fenómeno de los autistas sabios proporciona un fuerte apoyo para la existencia de inteligencias independientes, incluyendo la inteligencia musical.[†] Un **autista sabio** es un individuo con retraso mental que tiene un talento sumamente desarrollado en una sola área, como arte, cálculo rápido, memoria o música. Un ejemplo es el extraordinario caso de Leslie Lemke, que nació ciego, con retraso mental y parálisis cerebral; no se suponía que sobreviviera; su madre adoptiva tuvo que persuadirlo de que tomara leche de un biberón; posteriormente, lo fijó con correas a su propia espalda para ayudarlo a aprender a caminar. A pesar de sus graves discapacidades, Leslie se enamoró del piano y mostró una increíble precocidad para sacar melodías en él. Después de unos cuantos años, a la edad de 18, podía escuchar una pieza clásica de piano una sola vez y después tocarla sin cometer ningún error (Patton, Payne y Beirne-Smith, 1986). El lector puede encontrar estudios de caso adicionales sobre autistas sabios en Miller (1989) y Treffert (1989).

Recientemente, Gardner (1998) ha añadido tres candidatas tentativas a su lista de inteligencias. Éstas son naturalista, espiritual y existencial. La inteligencia naturalista es del tipo mostrado por personas que pueden discernir patrones dentro de la naturaleza. Charles Darwin sería el principal ejemplo de una persona de ese tipo. Gardner considera que la evidencia para este tipo de inteligencia es relativamente fuerte. En contraste, existen pruebas menores de que la inteligencia espiritual (una preocupación acerca de las cuestiones cósmicas y espirituales en el propio desarrollo) y la inteligencia existencial (una preocupación con cuestiones fundamentales, incluyendo el significado de la vida) sean inteligencias independientes. En general, la teoría de las inteligencias múltiples es convincente por su sencillez, pero existe poca investigación empírica acerca de su validez.

STERNBERG Y LA TEORÍA TRIÁRQUICA DE LA INTELIGENCIA

Sternberg (1985b, 1986, 1996) asume una perspectiva mucho más amplia acerca de la naturaleza de la inteligencia que la mayoría de los teóricos anteriores. Además de proponer que se requieren ciertos mecanismos mentales para la conducta inteligente, también enfatiza que la inteligencia implica adaptación al ambiente de la vida real. Su teoría enfatiza lo que él denomina inteligencia exitosa o "la capacidad para adaptarse, moldear y seleccionar los ambientes para lograr las propias metas y aquéllas de la propia sociedad y cultura" (Sternberg y Kaufman, 1998, p. 494).

La teoría de Sternberg se denomina **triárquica** (gobernada por tres) porque trata sobre tres aspectos de la inteligencia: componencial, basada en la experiencia y contextual. Cada uno de estos tipos de inteligencia tiene dos o más subcomponentes. La teoría completa se resume en el cuadro 5–3.

La **inteligencia componencial** consta de los mecanismos mentales internos responsables de la conducta inteligente. Los componentes de la inteligencia satisfacen tres funciones diferentes. Los **metacomponentes** son los procesos ejecutivos que dirigen las actividades de todos los demás componentes de la inteligencia; son responsables de determinar la naturaleza de un problema intelectual, seleccionando una estrategia para resolverlo y asegurándose de que se lleve a cabo la tarea. Los metacomponentes reciben realimentación constante acerca de cómo van las cosas en la solución del problema. Las

[†] Históricamente, los autistas sabios se han llamado **idiotas sabios**, que se refiere en sentido literal a una persona que al mismo tiempo tiene un retraso profundo y, sin embargo, es un "sabio". Por razones obvias, se ha hecho a un lado el primer término.

Cuadro 5–3. Resumen de la teoría triárquica de la inteligencia de Sternberg

Inteligencia componencial
- Metacomponentes o procesos ejecutivos (p. ej., planificación).
- Componentes de ejecución (p. ej., razonamiento silogístico).
- Componentes de adquisición de conocimiento (p. ej., capacidad para adquirir palabras de vocabulario).

Inteligencia basada en la experiencia
- Capacidad para enfrentar la novedad.
- Capacidad para automatizar el procesamiento de información.

Inteligencia contextual
- Adaptación al ambiente de la vida real.
- Selección de un ambiente adecuado.
- Moldeamiento del ambiente.

Fuente: Resumido de Sternberg, R. J. (1986). *Intelligence applied: Understanding and increasing your intellectual skills.* San Diego, CA: Harcourt Brace Jovanovich.

personas con una gran capacidad en el aspecto metacomponencial de la inteligencia tienen muy buenos resultados para asignar sus recursos intelectuales.

En un estudio sobre solución de problemas en el que se emplearon formas novedosas de analogías, Sternberg (1981) encontró que la inteligencia superior se relaciona con la utilización de un tiempo relativamente mayor en la planificación global o de orden superior y relativamente menos tiempo en la planificación local o de orden inferior. Por ejemplo, considérese este problema de analogía:

Hombre:Piel: (Perro, Árbol): (Corteza, Gato)

El examinando debe elegir los dos términos correctos de la derecha que completarán la analogía (las opciones correctas son árbol y corteza). Al utilizar medidas de tiempo de reacción para una serie de dichos problemas novedosos o no consolidados, Sternberg (1981) encontró que las personas con mayor inteligencia pasan más tiempo en la planificación global —formación de una macroestrategia que se aplique a éste y otros problemas similares— que aquéllas con inteligencia menor. Así, un aspecto importante de la inteligencia consiste en conocer cuándo hay que dar marcha atrás y asignar el esfuerzo intelectual en lugar de atacar de manera obtusa un problema difícil.

Los componentes de **ejecución** son los procesos mentales bien establecidos que podrían utilizarse para llevar a cabo una tarea o resolver un problema. Estos aspectos de la inteligencia son los que con toda probabilidad se miden mejor a través de las pruebas existentes de la misma. Ejemplos de los componentes de ejecución incluyen memoria a corto plazo y razonamiento silogístico.

Los componentes de **adquisición del conocimiento** son los procesos que se utilizan en el aprendizaje. Sternberg ha enfatizado que a fin de entender aquello que hace que algunas personas sean más hábiles que otras, debe comprenderse su capacidad cada vez mayor para adquirir desde un principio dichas habilidades. Un ejemplo es el conocimiento de vocabulario que se aprende principalmente en contexto más que a través de instrucción directa. Las personas más inteligentes están mejor capacitadas para utilizar los contextos circundantes para descubrir lo que significa una palabra; es decir, tienen mayores habilidades para la adquisición de conocimiento. Su mayor vocabulario se debe, en gran medida, a su mayor capacidad para "absorber" los significados de las palabras que ven y escuchan en su ambiente. Así, el vocabulario es una excelente medida de la inteligencia porque refleja la capacidad de las personas para adquirir información dentro de un contexto.

El segundo aspecto de la teoría de Sternberg incluye la **inteligencia basada en la experiencia**. Según la teoría, una persona con buena inteligencia basada en la experiencia puede enfrentarse de manera eficaz a tareas novedosas. Este aspecto de su teoría explica por qué Sternberg es tan crítico con la mayor parte de las pruebas de inteligencia. En la generalidad de los casos, las pruebas existentes miden aquello que ya se ha aprendido, al presentarle al individuo tareas que ya conoce. Según Sternberg,

la inteligencia también implica la capacidad para aprender y pensar dentro de nuevos sistemas conceptuales, no sólo enfrentarse con tareas conocidas de antemano. Un segundo aspecto de este tipo de inteligencia es la capacidad para automatizar o "hacer rutinarias" las tareas que se enfrentan de manera repetida. Un ejemplo de la automatización que se aplica a la mayoría de las personas es la lectura, la cual se lleva a cabo en gran medida sin influencia del pensamiento consciente, pero cualquier tarea o habilidad mental se puede automatizar si se practica lo suficiente. Tocar música es un ejemplo de una habilidad de nivel extremadamente elevado que puede automatizarse con la práctica suficiente.

El tercer aspecto de la teoría de Sternberg incluye la **inteligencia contextual**. Ésta se define como "la actividad mental que participa en la adaptación propositiva, el moldeamiento y la selección de los ambientes reales idóneos para la propia vida" (Sternberg, 1986, p. 33). Este aspecto de la teoría parece reconocer que la conducta humana se ha moldeado a través de presiones selectivas durante la historia evolutiva. La inteligencia contextual tiene tres partes: adaptación, selección y moldeamiento.

La **adaptación** se refiere al desarrollo de habilidades que se requieren en el propio ambiente. La adaptación exitosa difiere de una cultura a otra. En la cultura de los pigmeos de África, la adaptación podría implicar la capacidad para rastrear elefantes y matarlos con flechas envenenadas. En las naciones industriales de occidente, la adaptación podría requerir la presentación favorable de uno mismo en una entrevista de trabajo.

La **selección** podría llamarse localización del nicho. Este aspecto de la inteligencia contextual requiere de la capacidad para dejar el ambiente en el que uno se encuentra y seleccionar uno diferente, más adecuado para los talentos y necesidades. Feldman (1982) ha ilustrado la manera en que la selección puede operar en la elección de carrera de los niños dotados intelectualmente, lo cual determina si alcanzan un éxito notable en la edad adulta. Esta investigadora hizo un seguimiento de los niños que aparecieron en los programas de radio y televisión *Quiz Kids* durante el decenio de 1950. Estos niños eran extremadamente inteligentes de acuerdo con normas convencionales, la mayoría con un CI de 140 o mayor; algunos lograron ser muy exitosos al llegar a la edad adulta. No obstante, la mayoría llevaron vidas bastante comunes y corrientes, desprovistas de los logros espectaculares que podrían haberse pronosticado con base en su precocidad. Aquellos que tuvieron mayor éxito habían encontrado ocupaciones muy adecuadas a sus capacidades e intereses. En resumen, habían seleccionado nichos ambientales que se adaptaban bien a su vida. Sternberg afirmaría que la capacidad para seleccionar dichos ambientes es un aspecto importante de la inteligencia.

El **moldeamiento** es otro modo de mejorar la adaptación entre uno mismo y el ambiente, en especial cuando la selección de un nuevo ambiente no es práctica. En esta aplicación de la inteligencia contextual, se moldea al ambiente de modo que se adapte mejor a las propias necesidades. Un empleado que convence a su jefe de hacer las cosas de manera diferente, ha utilizado el moldeamiento para hacer que su ambiente laboral sea más adecuado para sus propios talentos.

Aunque la teoría triárquica de Sternberg es el modelo más amplio y ambicioso que se haya propuesto a la fecha, no todos los investigadores en psicometría lo han adoptado. Detterman (1984) advierte que deberían investigarse los componentes cognitivos básicos de la inteligencia antes de introducir constructos de orden superior que pueden ser innecesarios. Rogoff (1984) cuestiona si las tres subteorías (componencial, basada en la experiencia y contextual) están suficientemente vinculadas. Pueden encontrarse otros comentarios acerca de la teoría triárquica en *Behavioral and Brain Sciences* (1984, pp. 287-304).

Independientemente del veredicto final acerca de la teoría triárquica de la inteligencia, la insistencia de Sternberg de que la inteligencia tiene varios componentes que no miden las pruebas tradicionales parece correcta para cualquiera que haya estudiado o aplicado dichas pruebas. Este autor cita el caso de un colega al

que se le pidió que examinara a varios residentes de una institución para individuos con retraso mental. Los residentes en cuestión habían planeado y ejecutado con éxito un escape de esta escuela, que era muy concienzuda en materia de seguridad, lo cual representó una proeza que requirió altos niveles de inteligencia práctica. Sin embargo, cuando se les aplicó el Test de Laberintos de Porteus (Porteus, 1965), una prueba estandarizada que supuestamente mide la capacidad de planificación, estos individuos no pudieron resolver de manera correcta ni siquiera el laberinto más sencillo. Sternberg (1986) ha dejado claro que la inteligencia simplemente tiene demasiados componentes como para medirla con cualquier prueba individual.

RESUMEN

1. A pesar de las conferencias y análisis eruditos, el concepto de "inteligencia" ha esquivado la definición consensual. No obstante, dos temas se repiten con cierta frecuencia en las definiciones de expertos acerca de la inteligencia. Según éstos, la inteligencia engloba 1) la capacidad para aprender de la experiencia y 2) la capacidad para adaptarse al propio ambiente.
2. Los conceptos de legos y expertos acerca de la inteligencia son muy semejantes. En orden de importancia, los legos consideran a la capacidad de solución de problemas prácticos, la capacidad verbal y la competencia social como los ingredientes fundamentales; los expertos piensan que la inteligencia verbal, la capacidad de solución de problemas y la inteligencia práctica son importantes.
3. Las primeras teorías acerca de la inteligencia, propuestas a finales del siglo XX, enfatizaban la agudeza sensorial. Tanto Sir Francis Galton como J. McKeen Cattell consideraban que la inteligencia se sustenta en capacidades sensoriales agudas. Ambos desarrollaron varias medidas sensoriales en sus fallidos intentos por medir la inteligencia.
4. A inicios del siglo XX, Charles Spearman propuso que la inteligencia constaba de dos tipos de factores; un solo factor general, *g*, y numerosos factores específicos, s_1, s_2, s_3, y así sucesivamente. Spearman ayudó a inventar el análisis factorial como auxiliar para sus investigaciones acerca de la naturaleza de la inteligencia.
5. L. L. Thurstone favoreció la perspectiva de que la inteligencia consta de aproximadamente siete factores grupales en vez de un solo factor general. Éstos eran comprensión verbal, fluidez de palabra, número, espacio, memoria asociativa, velocidad perceptual y razonamiento inductivo. En última instancia, Thurstone reconoció la existencia de *g* como factor de primer orden.
6. Raymond Cattell propuso que la inteligencia está formada por dos factores principales: inteligencia fluida (g_f) e inteligencia cristalizada (g_c). La primera es no verbal en su mayor parte y es una forma de eficiencia mental relativamente independiente de la cultura. La inteligencia cristalizada tiene una fuerte dependencia de la cultura y se emplea para tareas que requieren de una respuesta aprendida o habitual.
7. Jean Piaget propuso una tesis del desarrollo en su teoría acerca de la inteligencia. Sugirió que los esquemas —patrones organizados de conducta o estructuras mentales que conducen al conocimiento de cómo hacer las cosas— evolucionan hacia la maduración cada vez mayor a través de un proceso denominado equilibración.
8. En la teoría de Piaget, la asimilación es la aplicación de un esquema a un objeto, persona o acontecimiento. Si un esquema funciona, surge un estado de equilibrio, de lo

contrario el resultado es el desequilibrio —un estado de tensión dinámica. En este último caso, la persona debe adaptar el esquema para que funcione —un proceso denominado acomodación.

9. J. P. Guilford propuso un modelo de estructura intelectual (EI) para resumir sus puntos de vista acerca de la naturaleza multifacética de la inteligencia. Clasificó las capacidades intelectuales en tres dimensiones llamadas operaciones (cinco tipos), contenidos (cinco tipos) y productos (seis tipos). De esta manera, Guilford propuso en total 150 tipos diferentes de inteligencia.
10. Según la teoría del procesamiento simultáneo y sucesivo, el cerebro humano tiene dos formas específicas de procesamiento de información: simultánea, en la que se procesan a un mismo tiempo grupos de información principalmente espacial, y sucesiva, en la que la información se organiza de manera temporal en una serie lineal.
11. Algunos investigadores han estudiado los fundamentos biológicos de la inteligencia. Por ejemplo, varios estudios indican que la inteligencia psicométrica se correlaciona con aspectos de los patrones de ondas cerebrales. En algunos de ellos, la complejidad de la onda cerebral evocada (el potencial evocado promedio o PEP) se correlaciona a niveles de .70 a .80 con el CI medido.
12. H. Gardner propuso una teoría de las inteligencias múltiples basada de manera aproximada en el estudio de las relaciones cerebro-conducta. Afirma la existencia de varias inteligencias relativamente independientes que incluyen la inteligencia lingüística, musical, lógico-matemática, espacial, corporal-cinestésica y personal.
13. R. Sternberg propone una teoría triárquica de la inteligencia con los siguientes aspectos: inteligencia componencial (los mecanismos internos responsables de la conducta inteligente); inteligencia basada en la experiencia (la capacidad de manejar de manera eficaz las tareas novedosas) e inteligencia contextual (adaptación, moldeamiento y selección de ambientes de la vida real).

TÉRMINOS Y CONCEPTOS CLAVE

Definiciones operacionales p. 170
Definición real p. 170
Factor general p. 175
Factores específicos p. 175
Habilidades mentales primarias p. 177
Inteligencia fluida p. 178
Inteligencia cristalizada p. 178
Esquemas p. 179
Conservación p. 179
Equilibración p. 179
Asimilación p. 179
Acomodación p. 179
Producción divergente p. 182
Procesamiento simultáneo p. 182
Procesamiento sucesivo p. 182
Autista sabio p. 186
Inteligencia componencial p. 186
Inteligencia basada en la experiencia p. 187
Inteligencia contextual p. 188

TEMA 5B
Evaluación de las capacidades de lactantes y preescolares

Evaluación de la capacidad del lactante
Evaluación de la inteligencia del preescolar
Utilidad práctica de la evaluación con lactantes y preescolares
Resumen

El periodo de la lactancia y edad preescolar se extiende aproximadamente desde el nacimiento hasta los seis años de edad. Los cambios que ocurren durante este periodo son, obviamente, profundos. El lactante desarrolla los reflejos básicos, domina los hitos del desarrollo (asir, gatear, sentarse, ponerse de pie y demás), aprende un lenguaje y establece la capacidad para el pensamiento simbólico. Para la mayoría de los niños, el patrón y progreso del desarrollo se encuentra de manera visible dentro de límites normales.

No obstante, en ocasiones los padres y profesionales entrenados en la evaluación de lactantes y preescolares se encuentran con niños cuyo desarrollo parece lento, demorado o, incluso, que muestran un retraso evidente. Estos niños provocan una conmoción de preguntas ansiosas: ¿Qué tan retrasado está? ¿Cuáles son los prospectos de funcionamiento normal en la escuela? ¿Este niño logrará la independencia personal cuando llegue a la edad adulta?

En el extremo opuesto se encuentran aquellos niños precoces que alcanzan los hitos del desarrollo meses o años antes de lo programado en términos normativos. En estos casos, los orgullosos padres tienen un conjunto diferente de preocupaciones: ¿Qué tan avanzado está mi

hijo? ¿Cuáles son las áreas más fuertes y más débiles de funcionamiento intelectual? ¿Este niño será un adulto dotado?

Los instrumentos para la evaluación de lactantes y preescolares pueden ayudar a responder las preguntas acerca de niños en ambas situaciones. Por supuesto, estas pruebas también proporcionan información útil acerca de la mayoría de los niños que caen en medio de la distribución. En este tema se revisa la naturaleza y aplicación de las principales medidas para lactantes y preescolares. Estas herramientas incluyen pruebas individuales, inventarios del desarrollo y escalas de clasificación. Se comienza con una descripción de varios instrumentos destacados y después se investiga la pregunta fundamental del propósito o utilidad. ¿Cuál es el empleo de estas medidas? ¿Cuál es el significado de una puntuación en un inventario del desarrollo o en una prueba de inteligencia para preescolares? ¿Hasta qué grado estos procedimientos permiten el pronóstico de las capacidades adultas o, en ese caso, ayudan a vaticinar el desempeño inicial en la escuela? Estas preguntas tendrán más significado si se revisan primero los instrumentos relacionados.

La revisión se divide en dos partes: medidas para lactantes (dirigidas a niños desde el nacimiento hasta los dos años y medio de edad) y pruebas para preescolares (dirigidas a niños desde los 2 años y medio hasta los 6 años de edad). La división es un tanto arbitraria, pero no por completo. Las pruebas para lactantes tienden a ser multidimensionales y a tener cargas significativas en desarrollo sensorial y motor. A partir de los dos años y medio de edad, es típico que se utilicen las medidas estandarizadas como la Escala de Inteligencia Stanford-Binet: Cuarta Edición, la Batería de Evaluación Kaufman para Niños, las *Differential Ability Scales* (Escalas de Capacidad Diferencial) y las Escalas McCarthy de Aptitudes y Psicomotrocidad para niños para la evaluación de niños preescolares. Estas pruebas tienen fuertes cargas en habilidades cognitivas, como comprensión verbal y pensamiento espacial. Así, las escalas para lactantes y las pruebas para preescolares miden componentes un tanto diferentes de la capacidad intelectual.

EVALUACIÓN DE LA CAPACIDAD DEL LACTANTE

INVENTARIOS GESELL DEL DESARROLLO

Diseñados para medir el progreso en el desarrollo de lactantes y niños desde las cuatro semanas hasta los 60 meses de edad, los Inventarios Gesell del Desarrollo se publicaron por primera vez en 1925, y después se han revisado de manera periódica (Gesell, Ilg y Ames, 1974; Knobloch, Stevens y Malone, 1987). Casi todas las pruebas para lactantes han tomado prestados o han adaptado los reactivos de los inventarios originales diseñados por Arnold Gesell (1880-1961), de modo que es justo y apropiado que se comience la presente revisión con este instrumento.

Los Inventarios Gesell del Desarrollo proporcionan un procedimiento estandarizado para la observación y valoración de los logros en el desarrollo que tienen los niños en cinco áreas: motora gruesa, motora fina, desarrollo del lenguaje, conducta adaptativa y conductas personales-sociales. La mayor parte de los 144 reactivos en el inventario son puramente observacionales y se basan en la inspección directa de las respuestas de los niños a juguetes y a situaciones estándar. Por ejemplo, a continuación se presentan algunos reactivos ilustrativos que por lo común superaría un lactante de 40 semanas de edad.

Adaptativa
- Señala a una bolita en un vaso.
- Tira de un cordel para obtener un anillo.

Motora gruesa
- "Transita" por una barra utilizando dos manos.
- Se deja caer al piso con control.

Motora fina
- Toma una bolita de manera decidida.
- Utiliza asimiento de "tijeras" en un cordel.

Lenguaje
- Usa "da da" con significado.
- Responde a la palabra "no no".

Personal-social

Extiende un juguete, pero no lo suelta.

Pasa los brazos por el vestido, si se le ayuda a iniciar el movimiento.

El rango de edades del Inventario Gesell del Desarrollo es del nacimiento a los 60 meses. El genio de Gesell se encuentra en haber identificado situaciones que ocurren de manera natural en el hogar o la clínica y en utilizar objetos o tareas con enorme atractivo para lactantes y preescolares. En algunos casos, es necesaria la información de uno de los padres o de la persona encargada del cuidado del niño para calificar los reactivos individuales. A pesar del ambiente naturalista de la prueba, los observadores, con el entrenamiento adecuado, pueden lograr confiabilidades interobservadores a la mitad del rango de .90 (Knobloch, Stevens y Malone, 1987).

Los pediatras y otros especialistas en niños son quienes emplean principalmente los Inventarios Gesell del Desarrollo para identificar a lactantes y niños en riesgo de presentar alteración neurológica y retraso mental. Gesell nunca tuvo la intención de que su inventario fuera una prueba de inteligencia. Tenía una fuerte orientación biológica en su investigación y suponía que el desarrollo normal era una evolución madurativa que ocurría en una secuencia fácil de predecir. Gesell determinó que el desarrollo normal es un fenómeno con límites temporales: la variabilidad de edades para el logro de los hitos del desarrollo en la lactancia es generalmente pequeña, en el rango de unas cuantas semanas para muchas tareas. Por tanto, la demora grave para alcanzar a los hitos del desarrollo, establecidos por Gesell de manera concienzudamente cronológica, puede indicar alteración neurológica o retraso mental (Honzik, 1983; Lewis y Sullivan, 1985). Varios estudios indican que los Inventarios Gesell del Desarrollo funcionan bien en la detección de lactantes en riesgo (Knobloch, Stevens y Malone, 1987).

Aunque los Inventarios Gesell del Desarrollo se emplean principalmente para la detección y diagnóstico clínicos, Knobloch, Stevens y Malone (1987) proporcionan una base más o menos definida para la obtención de Cocientes del Desarrollo para las cinco áreas y el desarrollo general. La fórmula es la siguiente:

$$CD = \frac{\text{Edad de maduración}}{\text{Edad cronológica}} \times 100$$

La edad de maduración se basa en la "imagen clínica total" de los hitos del desarrollo superados y no superados en cada área. Aunque no se proporcionan criterios precisos, la edad de maduración para un lactante parece ser la edad del desarrollo en la que se pasan la mayor parte de los reactivos. Dado que sus propiedades técnicas no se han estudiado de manera adecuada, el Cociente del Desarrollo debe utilizarse principalmente como herramienta de investigación.

Las pruebas Gesell son muy respetadas debido a que proporcionan descripciones detalladas de los hitos del desarrollo en la lactancia que no tienen paralelo en la bibliografía sobre evaluación infantil (Nuttall, Romero y Kalesnik, 1992). Sin embargo, el empleo de los inventarios Gesell como instrumento psicométrico ha recibido agudas críticas en años recientes. El problema básico parece residir en una falta de atención a los criterios formales de confiabilidad y validez. Por ejemplo, en los primeros manuales de Gesell los informes sobre la confiabilidad *test-retest* son mínimos o nulos. Cuando los investigadores contemporáneos examinaron esta propiedad de las pruebas Gesell, los resultados fueron sorprendentes. Lichtenstein (1990) informó de una correlación *test-retest* de sólo .73 con una muestra de 46 niños, lo cual cae muy por debajo del nivel recomendado de .90 para la toma de decisiones sobre individuos (Nunnally, 1978; Salvia & Ysseldyke, 1991). Banerji (1992) concluyó que las pruebas Gesell funcionaban de manera deficiente como instrumentos de detección de la preparación para la escuela. En general, los especialistas en educación se muestran cautelosos en cuanto a la utilización de estas pruebas para la toma de decisiones sobre asignación o retención escolar.

ESCALAS ORDINALES DE DESARROLLO PSICOLÓGICO

Las *Ordinal Scales of Psychological Development* (Escalas Ordinales de Desarrollo Psicológico), llamadas de aquí en adelante Escalas Ordinales, se diseñaron como una herramienta basada en la teoría de Piaget para la medición del desarrollo intelectual entre niños de 2 semanas y 2 años de edad (Uzgiris y Hunt, 1989). Dichas heramientas incluyen seis escalas, cada una diseñada para medir una capacidad específica que surge durante el primer estadio piagetiano de la inteligencia sensoriomotora. Cada escala consta de 5 a 15 pasos ordinales independientes; es decir, los reactivos se ordenaron en una secuencia del desarrollo que normalmente no varía.

Las escalas son las siguientes:

- Exploración visual y permanencia de objeto.
- Desarrollo de medios-fines.
- Imitación vocal y de ademanes.
- Desarrollo de la causalidad operacional.
- Construcción de relaciones de objeto en el espacio.
- Desarrollo de esquemas para relacionarse con los objetos.

Ante las muchas adversidades que surgen cuando se examina a los lactantes —es posible que lloren, regurgiten, se alejen gateando, ignoren la tarea, se duerman o se queden fijados en la barba del examinador— las escalas de este instrumento poseen fuertes propiedades psicométricas. En un estudio con 84 lactantes, las Escalas Ordinales mostraron excelente confiabilidad interobservadores (media de 96%), buena consistencia *test-retest*, ordinalidad respetable y correlaciones muy fuertes con la edad (Uzgiris, 1976). En pocas palabras, este instrumento parece ser un índice psicométricamente sólido de la inteligencia sensoriomotora.

Uzgiris (1983) considera que el funcionamiento intelectual en la lactancia es cualitativamente diferente y "necesita comprenderse por sus características propias". Las Escalas Ordinales se desarrollaron como un medio para investigar la inteligencia del lactante dentro del esquema teórico de Piaget; por esta razón, Uzgiris no pretende que su instrumento sirva como medio de predicción. En general, las correlaciones entre las puntuaciones de la escala y el CI posterior son muy bajas hasta que los lactantes tienen cuando menos 18 meses de edad. Muy pocos clínicos utilizan el instrumento para la detección del desarrollo; sin embargo, Dunst (1980) ha abogado por el empleo de estas escalas como base para el diseño de un currículo sólido en términos del desarrollo para niños con discapacidad. Recientemente, Auer y Reisberg (1996) han hecho surgir la fascinante posibilidad de que las Escalas Ordinales puedan emplearse para la evaluación cognitiva de la demencia grave en ancianos.

ESCALAS BAYLEY DE DESARROLLO INFANTIL – II

Después de decenios de preponderancia en el campo de la evaluación de lactantes, las Escalas Bayley de Desarrollo Infantil han tenido una revisión reciente (Bayley, 1969, 1993). El formato de la escala es el mismo —la Escala Mental y la Escala Motora proporcionan puntuaciones cuantitativas estándar normalizadas, con media de 100 y desviación estándar de 16— pero las *Bayley Scales of Infant Development–II* cubren un rango más amplio de edades, que va desde 1 hasta 42 meses. El tercer componente, la Escala de Calificación Conductual, consta de 30 reactivos diseñados para evaluar atención, orientación, regulación emocional y calidad motora. La Bayley–II tiene nuevas normas obtenidas con una muestra aleatoria estratificada de 1 700 niños que se asemeja mucho a los datos estadísticos del censo de EUA para 1988 en cuanto a edad, sexo, origen étnico, región y educación de los padres.

La Escala Mental mide las siguientes capacidades:

- Agudeza sensorial/perceptual.
- Adquisición de constancia objetal.
- Memoria, aprendizaje y solución de problemas.
- Vocalización, comunicación verbal.

- Evidencia inicial de pensamiento abstracto.
- Habituación.
- Mapeo mental.
- Lenguaje complejo.
- Formación de conceptos matemáticos.

La Escala Motora evalúa las siguientes habilidades:

- Grado de control del cuerpo.
- Coordinación de músculos grandes.
- Control motor fino de manos y dedos.
- Movimiento dinámico.
- Praxis dinámica.
- Imitación postural.
- Estereognosis.

La calidad técnica y la excelente estandarización de las escalas Bayley señalan a esta prueba como el pináculo psicométrico en su campo (Sattler, 1988). Aunque la escala Bayley–II tiene sólo una modesta cantidad de investigación sobre validez, se asemeja en gran medida a su antecesora, para la cual se puede citar una enorme cantidad de evidencias de validación. Así, la validez de la Bayley–II descansa, en parte, en su semejanza con la escala Bayley original. En cuanto a validez, el manual de la escala Bayley informa de una correlación de .57 entre la Escala Mental y el CI de la Stanford-Binet para 120 niños con edades de 24 a 30 meses. Self y Horowitz (1979) revisaron la voluminosa literatura acerca de los correlatos de las puntuaciones de la escala Bayley. Esta escala muestra fuerte relación con la Stanford-Binet, con las Escalas Wechsler, el desempeño en tareas piagetianas, clase social y factores ambientales. También, las puntuaciones muy bajas en la escala Bayley predicen un resultado deficiente del desarrollo en la infancia tardía (VanderVeer y Schweid, 1974). Rhodes, Bailey y Yow (1983) citan evidencia adicional sobre validez.

Un estudio reciente sobre la validez de las *Bayley Scales of Infant Development–II* (BSID-II) con lactantes prematuros encontró una fuerte concordancia entre esta prueba y la primera, lo cual sustenta la validez clínica de la revisión (Goldstein, Fogle, Wieber y O'Shea, 1995). Un estudio con lactantes australianos sanos informó que las puntuaciones de la BSID-II eran apropiadamente más bajas que las de la BSID, lo cual indica que las normas para la primera edición en verdad eran obsoletas (Tasbihsazan, Nettelbeck y Kirby, 1997). A pesar de estos estudios en apoyo a su validez, Nellis y Gridley (1994) sugieren precaución con la BSID-II hasta obtener mayor información.

Las escalas Bayley requieren mayor habilidad y tiempo (45 a 75 minutos) para aplicar e interpretar la prueba que instrumentos comparables como la Denver-2. En consecuencia, se reservan para evaluaciones especiales y para aplicaciones de investigación; por lo común no se les utiliza como instrumentos rutinarios de detección.

RESUMEN: MEDIDAS ADICIONALES DE LA CAPACIDAD EN LA LACTANCIA

La evaluación de los lactantes es muy importante y, sin embargo, muy difícil. Por lo común, este grupo no sigue las instrucciones y es posible que no pueda verbalizar lo que sabe. La evaluación de las capacidades de los lactantes representa un reto extraordinario. No obstante, docenas de autores de pruebas han respondido al llamado. Inclusive una revisión breve de los instrumentos alternativos abarcaría un capítulo completo. En el cuadro 5–4 se proporciona un breve resumen de los enfoques más conocidos. En la mayor parte de estos instrumentos se emplea la observación o la presentación de tareas sencillas al examinando. Para reseñas adicionales sobre evaluación de lactantes, se motiva al lector a revisar los trabajos de Nuttall, Romero y Kalesnik (1992), Ricciuti (1994) y Salvia e Ysseldyke (1991).

EVALUACIÓN DE LA INTELIGENCIA DEL PREESCOLAR

Los niños preescolares exhiben gran variabilidad en cuanto a madurez emocional y receptividad hacia los adultos. Un niño puede animarse ante el examinador y luchar por dar su mejor

Cuadro 5–4. Medidas adicionales de la capacidad en la lactancia

Inventario del Desarrollo (Battelle Developmental Inventory, BDI; Newborg, Stock, Wnek, Guidubaldi y Svinicki, 1984). Del nacimiento a los ocho años de edad; los 341 reactivos evalúan los dominios Personal-Social, Adaptativo; Motor; Comunicación; Cognitivo y Total. La batería completa requiere de 1 a 2 horas para su aplicación; una versión del BDI para detección (96 reactivos) ha recibido graves críticas.

Developmental Assessment of Young Children (Evaluación del Desarrollo de Niños Pequeños DAYS; Voress y Maddox, 1998). Del nacimiento a los seis años de edad; la evaluación en cinco dominios (cognición, comunicación, socioemocional, físico y adaptativo) se realiza a través de observación, entrevistas con los encargados del cuidado del niño y evaluación directa. La DAYS proporciona una evaluación breve (20 minutos) con base en datos normativos sobresalientes (1 300 niños divididos en 23 grupos por edad que se aproximan a los datos del censo de 1996). Los cinco índices resultantes y el índice global son sumamente confiables (con coeficientes que van de .90 a .99).

Developmental Indicators for the Assessment of Learning-3 (Indicadores del Desarrollo para la Evaluación del Aprendizaje–3 DIAL-3; Mardell-Czudnowski y Goldenberg, 1998). De los 3 a los 6 años de edad; los dominios que se evalúan incluyen Motor (p. ej., atrapar, cortar, escribir), Conceptos (p. ej., nombrar, contar, clasificar) y Lenguaje (p. ej., sustantivos/verbos, solución de problemas, extensión de la oración). La confiabilidad *test-retest* en la parte alta del rango de .80 es extraordinaria para un instrumento de este tipo. En el mismo estuche se tienen disponibles versiones en inglés y español.

Early Screening Inventory (Inventario de Detección Temprana, ESI; Meisels, Wiske y Tivnan, 1984). Edades de 3 a 6 años; es un instrumento breve de detección que proporciona puntuaciones en cuatro áreas —Dibujo de la Figura Humana, Visomotora/Adaptativa; Lenguaje y Cognición y Motora Gruesa/Consciencia Corporal. La puntuación total se utiliza para clasificar a los niños en uno de tres grupos de canalización: "Bien" (de superior al promedio a menos 1 *DE*); "reevaluar" (entre menos 1 y menos 2 *DE*) y "canalizar" (inferior a menos 2 *DE*).

Early Screening Profiles (Perfiles de Detección Temprana, ESP; Harrison, Kaufman, Kaufman y col., 1990). Edades de 2 a 6 años; los dominios que se evalúan incluyen Cognitivo/Lenguaje, Motor y Autoayuda/Social; cuatro Exámenes (Articulación, Conducta, Historial de salud y Hogar) complementan la evaluación. Este instrumento tiene fuertes cualidades psicométricas; el manual proporciona información detallada en siete estudios de validación llevados a cabo de manera independiente con respecto al estudio de estandarización. Barnett (1995) ofrece una revisión escéptica; Telzrow (1995) se muestra más positivo.

Gesell Child Development Age Scale (Escala Gesell de Edad del Desarrollo Infantil, GCDAS; Cassell, 1990). De los 18 meses a los 10 años de edad; esta prueba intenta poner en términos operacionales la teoría de etapas del desarrollo infantil de Gesell al pedirle a la madre, maestro o clínico que respondan a 100 reactivos de falso-verdadero, apropiados para cada edad, dentro de un conjunto más amplio de 240 reactivos. Se pueden utilizar hasta tres calificadores para valorar a un niño; los resultados incluyen una gráfica de la edad cronológica contra la edad del desarrollo en 10 áreas del desarrollo. La GCDAS es una prueba prometedora que requiere de mayor investigación en cuanto a cualidades psicométricas (Lang, 1995).

desempeño en todas las preguntas. Otro puede mirar fijamente al piso y quedarse mudo mientras intenta realizar una sencilla tarea de diseño con cubos. En el primer caso se puede asegurar que los resultados de la prueba son un índice apropiado del funcionamiento cognitivo; pero en el segundo, prevalece la incertidumbre. ¿La falta de receptividad señala una carencia de habilidad o de cooperación? Con niños preescolares, el examinador requiere de una gran humildad. Scarr (1981) ha expresado este sentimiento de la siguiente manera:

Cuando se mide el funcionamiento cognitivo de un niño, también se mide la cooperación, atención, persistencia, capacidad para estarse quieto y receptividad social ante una situación de evaluación.

El peligro especial en la evaluación preescolar es que el examinador puede inferir que una puntuación baja indica un bajo nivel de funcionamiento cognitivo cuando, en realidad, el niño tan sólo es incapaz de quedarse quieto, atender, cooperar y así sucesivamente. La evaluación preescolar requiere abordarse con un cuidado poco usual, a fin de evitar las consecuencias negativas de la etiquetación y el diagnóstico excesivo de padecimientos discapacitantes.

Existen varias pruebas de inteligencia de aplicación individual adecuadas para niños preescolares. Schakel (1986) ha llamado a las siguientes pruebas como "las cuatro grandes":

- Escala Wechsler de Inteligencia para los Niveles Preescolar y Primaria (WPPSI-R)
- Escala de Inteligencia Stanford-Binet: Cuarta Edición (SB:CE)
- Batería de Evaluación Kaufman para Niños (K-ABC)
- Escalas McCarthy de Aptitudes y Psicomotricidad para Niños (MSCA)

Éstas son las pruebas de inteligencia más comunes en niños preescolares. La última se acerca rápidamente a la obsolescencia (se publicó en 1972); a menos que se le revise, los psicólogos escolares pronto comenzarán a hablar de "las tres grandes". Por supuesto, algunos de estos instrumentos van más allá de la edad preescolar y llegan hasta los inicios de la niñez. La SB:CE se utiliza también con adultos. Se revisan éstas y otra prueba prometedora adicional:

- las *Differential Ability Scales*.

ESCALA WECHSLER DE INTELIGENCIA PARA LOS NIVELES PREESCOLAR Y PRIMARIA–REVISADA (WPPSI-R)

El WPPSI-R es muy similar a su antecesor, pero ofrece una actualización de normas y aplicación para un rango más amplio de edades —de los 3 a los 7 años, 3 meses (Wechsler, 1989). Además, se revisaron varios reactivos obsoletos y parciales y se añadió una versión de Composición de objetos a las 11 subpruebas originales. La confiabilidad y validez del WPPSI-R son muy similares a las de la versión anterior de esta prueba. Salvia e Ysseldyke (1991) resumen las confiabilidades de división en mitades como sigue: CI Verbal (.86 a .96), CI de Ejecución (.85 a .93) y CI Total (.90 a .97). Las confiabilidades de las subpruebas del WPPSI-R son sustancialmente más débiles; sus interpretaciones deben restringirse a las puntuaciones mixtas de CI. Las normas para el WPPSI-R se basan en una muestra cuidadosamente estratificada de 1 700 niños y en edad, sexo, región geográfica, origen étnico y educación y ocupación de los padres; su validez se basa, en parte, en su parecido con el WPPSI, que obtuvo grandes alabanzas de los revisores. Sattler (1988) revisó varias docenas de estudios en apoyo de la validez concurrente y predictiva del WPPSI, y concluyó que la prueba sirve como un excelente medio de predicción a largo plazo de la inteligencia y del desempeño escolar en la adolescencia.

La investigación inicial con el WPPSI-R confirma la validez predictiva de este instrumento para el desempeño escolar posterior. Por ejemplo, Kaplan (1996) determinó que los resultados preescolares del WPPSI-R predicen en gran medida las puntuaciones de rendimiento escolar de los niños desde el jardín de infantes hasta el tercer grado. Los resultados con niños en tercer grado se resumen en el cuadro 5–5 y revelan que los CI Verbal y Total constituyeron medios de predicción mucho más poderosos del rendimiento posterior que el CI de Ejecución.

Las subpruebas del WPPSI-R incluyen las siguientes:

Verbal	Ejecución
Información	Composición de objetos
Comprensión	Diseños geométricos
Aritmética	Diseño con cubos
Vocabulario	Laberintos
Semejanzas	Figuras incompletas
Frases	Casa de los animales

Cuadro 5–5. Correlaciones entre los CI preescolares del WPPSI-R y resultados posteriores en una prueba de rendimiento para 72 niños de tercer grado

	Puntuaciones del WPPSI-R		
Puntuaciones del programa completo de prueba–III	CIV	CIE	CIT
Capacidad verbal	.62*	.24	.52*
Comprensión auditiva	.46*	.02	.30
Comprensión de lectura	.48*	–.03	.28
Mecánica de la escritura	.54*	.16	.42
Proceso de escritura	.45*	.08	.32
Capacidad cuantitativa	.44	.17	.36
Matemáticas	.62*	.34*	.58*

*p < .005.

Fuente: Reimpreso con autorización de Kaplan, C. (1996). Predictive validity of the WPPSI-R: a four year follow-up study. *Psychology in the Schools, 33,* 211-219.

Tres de estas 12 subpruebas (Frases, Diseños geométricos y Casa de los animales) se encuentran sólo en el WPPSI-R y se resumen a continuación (las otras nueve subpruebas, comunes en todas las escalas Wechsler, se analizan en el Tema 6A: Pruebas individuales de inteligencia). Frases es una subprueba complementaria en el WPPSI-R. Ésta requiere que el niño repita al pie de la letra una oración que le ha leído en voz alta el examinador. El reactivo más fácil se encuentra al nivel de "Juan tiene un coche verde", mientras el más difícil es mucho más largo y consta de dos oraciones conectadas como éstas:

> "Este viernes visitaremos la huerta del granjero. Lleva una moneda para que puedas comprar una calabaza."

La subprueba de Diseños geométricos está formada por 10 diseños —incluyendo un círculo, un cuadrado y un rombo— que el niño debe copiar. Esta subprueba es una medida de las capacidades perceptuales y de organización visomotora. Por último, la subprueba de Casa de los animales requiere que el niño coloque un cilindro de un color específico (negro, blanco, azul, amarillo) en un orificio debajo del animal apropiado (perro, pollo, pez y gato, respectivamente). Hay 25 animales distribuidos en secuencia aleatoria en una matriz de 5 x 5. La puntuación inicial en esta subprueba es la cantidad de tiempo requerido para colocar un cilindro bajo cada animal. Los errores se restan de la puntuación total. El éxito en esta subprueba requiere capacidad de aprendizaje, destreza manual y atención sostenida durante los minutos que se necesitarían para colocar un cilindro apropiado bajo cada uno de los 25 animales.

La ampliación de la cobertura de edad hacia un límite inferior de tres años constituye una adición bien recibida para el WPPSI-R, dado que la identificación temprana de las dificultades del desarrollo es esencial para su tratamiento. También, las normas de CI del WPPSI-R se extendieron hacia abajo hasta llegar a una puntuación de 41, que se encuentra aproximadamente a 3.9 desviaciones estándar por debajo de la media de la población. En especial cuando se le utiliza junto con una evaluación de conducta de adaptación, el WPPSI-R es una herramienta esencial para el diagnóstico del retraso mental de leve a grave en niños preescolares y en edad escolar temprana. Las normas de CI para esta prueba también se extienden muy por arriba del rango necesario para la identificación de inteligencia dotada en la mayor parte de los entornos escolares. Estos aspectos han hecho que el WPPSI-R sea muy popular entre los psicólogos escolares y especialistas en desarrollo temprano.

ESCALA DE INTELIGENCIA STANFORD-BINET: CUARTA EDICIÓN

Con un rango de edades desde los dos años hasta la edad adulta, la Escala de Inteligencia Stanford-Binet: Cuarta Edición (SB:CE) es una de esas pruebas poco comunes diseñadas para utilizarse por igual con preescolares, niños y adultos (Thorndike, Hagen y Sattler, 1986). En el siguiente capítulo sobre pruebas individuales y grupales de inteligencia se presenta un análisis

detallado de las propiedades psicométricas de la SB:CE. Algunos comentarios en este capítulo resumirán su valor en la evaluación preescolar.

La SB:CE consta de 15 subpruebas, pero no todas se aplican a cada uno de los grupos por edad. En el cuadro 5–6 se muestran las subpruebas en general y aquellas que por lo común se aplican a niños preescolares (hasta los cinco años de edad). El lector observará que la SB:CE produce varias puntuaciones de subprueba, cuatro de área y una mixta total, que ya no se denomina CI. Por desgracia, en el caso de preescolares no se toma una muestra de la misma profundidad de las cuatro áreas de contenido. El área de Razonamiento verbal está bien representada con tres subpruebas, pero la puntuación de Cuantificación se basa en una sola subprueba. Las otras dos áreas (Razonamiento abstracto/visual y Memoria a corto plazo) se basan sólo en los resultados de dos subpruebas. Como se analiza en el siguiente capítulo, Sattler (1988) defiende una solución de dos factores para informar las puntuaciones de la SB:CE (Comprensión verbal y Razonamiento no verbal/Visualización) que con toda seguridad es el enfoque preferible en el caso de preescolares.

Un aspecto esencial de la SB:CE es que la puntuación mixta total es sumamente comparable con otros pilares de la evaluación preescolar, como el WPPSI-R y el WISC-III. Por ejemplo, el manual del WPPSI-R informa de puntuaciones globales similares para 115 niños, de 4 a 7 años de edad, examinados con ambos instrumentos: CI promedio en el WPPSI-R de 105.3 contra puntuación mixta de 107.2 en la SB:CE (Wechsler, 1989). En un estudio con 30 niños preescolares, el CI promedio en el WPPSI-R fue de 94.1, mientras la puntuación mixta de la SB:CE fue de 95.8 (McCrowell y Nagle, 1994). No obstante, los componentes verbales de las dos pruebas difieren de manera significativa: 95.5 para el CIV en el WPPSI-R y 101.6 para Razonamiento verbal en la SB:CE. Rust y Lindstrom (1996) encontraron puntuaciones comparables entre la SB:CE y el WISC-III con 57 voluntarios (de 6 a 17 años de edad), con puntuaciones totales que diferían, en promedio, en menos de dos puntos. Lavin (1996) también informó puntuaciones totales casi idénticas en estos dos instrumentos para 40 niños con edades de 6 a 16 años. Sin embargo, en el examen de niños con problemas identificados del desarrollo, Lukens y Hurrell (1996) encontraron que las puntuaciones en el WISC-III eran menores en 29 de 31 casos, lo cual indica que la SB:CE puede subdiagnosticar el retraso mental. Se puede encontrar un análisis amplio de la SB:CE en el contexto de la evaluación preescolar en Nuttall, Romero y Kalesnik (1992). Laurent, Swerdlik y Ryburn (1992) proporcionan una reseña de la investigación sobre validez con la SB:CE.

Cuadro 5–6. Subpruebas y áreas de la Escala de Inteligencia Stanford-Binet, cuarta edición

Razonamiento Verbal
Vocabulario*
Comprensión*
Absurdos*
Relaciones Verbales
Razonamiento Abstracto/Visual
Análisis de Patrones*
Copiado*
Matrices
Doblado y cortado de papel
Razonamiento Cuantitativo
Cuantificación*
Series de números
Construcción de ecuaciones
Memoria a Corto Plazo
Memoria de cuentas*
Memoria de oraciones*
Memoria de dígitos
Memoria de objetos

*Denota una subprueba que comúnmente se emplea con niños preescolares.

BATERÍA DE EVALUACIÓN KAUFMAN PARA NIÑOS (K-ABC)

La K-ABC es una medida combinada de inteligencia y rendimiento que se construyó de manera

aproximada dentro del esquema teórico de la neuropsicología moderna (Luria, 1966; Das, Kirby y Jarman, 1979). Muchas de las subpruebas de esta batería se parecen a pruebas neuropsicológicas, las cuales se analizarán con mayor detalle en un tema posterior; la K-ABC se orienta principalmente hacia la evaluación psicoeducativa y la planificación educativa (Kaufman y Kaufman, 1983); sus defensores afirman que posee mayor relevancia para la planificación psicoeducativa que pruebas tradicionales como las escalas Wechsler y la Stanford-Binet.

La K-ABC está diseñada para niños de 2 años y medio a 12 años y medio de edad y consta de 16 subpruebas, de las cuales no se aplican más de 13 a cada niño evaluado (figura 5–5). Diez de las subpruebas producen la Puntuación Mixta de Procesamiento Mental, que está normada para generar el promedio familiar de 100 y desviación estándar de 15. Las otras seis componen la Escala de Rendimiento. Las 10 subpruebas de procesamiento mental se descomponen en dos escalas globales: Escala de procesamiento simultáneo (siete subpruebas) y Escala de procesamiento secuencial (tres subpruebas).

Un objetivo de la K-ABC consiste en producir puntuaciones que se traduzcan en intervenciones educativas. Las escalas Secuencial y Simultánea se basan aproximadamente en conceptos neuropsicológicos y se supone que reflejan el estilo de solución de problemas y de procesamiento de información del niño. Las subpruebas de Procesamiento secuencial requieren de la disposición en serie o temporal del contenido verbal, numérico o visoperceptual. Se supone que los niños con puntuaciones altas en esta escala —aprendices secuenciales— aprenden mejor al enfrentar pequeñas cantidades de información en orden consecutivo, paso por paso, como en una serie de instrucciones verbales claras. En contraste, las subpruebas de Procesamiento simultáneo requieren que el niño sintetice y organice el contenido visoperceptual o espacial de manera inmediata o global. Se supone que los niños con puntuaciones altas en esta escala —aprendices simultáneos— aprenden mejor a través de la integración y síntesis

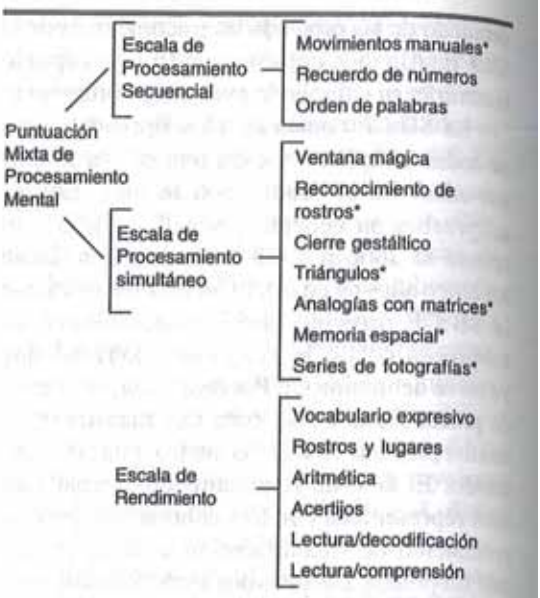


Figura 5–5. Subpruebas y escalas de la Batería de Evaluación Kaufman para niños.

de muchos trozos relacionados de información al mismo tiempo, como se encuentra en medios visuales (ilustraciones, mapas o gráficas).

Kaufman, Kaufman y Goldsmith (1984) proporcionan pautas y ejemplos para la enseñanza de la lectura, ortografía y aritmética a niños con base en las fortalezas secuenciales o simultáneas en la K-ABC. Aunque la teoría es convincente, la sustentación para la supuesta interacción entre aptitud y tratamiento de la K-ABC es dispar en el mejor de los casos. Por ejemplo, Fisher, Jenkins, Bancroft y Kraft (1988) equipararon estrategias de enseñanza con los estilos cognitivos secuencial/simultáneo (como se determinan a partir de la K-ABC) para 57 estudiantes de educación primaria inscritos en una clínica de discapacidad para el aprendizaje. Aunque, en general, los resultados prestaron apoyo a la interacción predicha entre aptitud y tratamiento, los efectos fueron pocos y no tuvieron ninguna importancia práctica.

Ambas escalas de la Puntuación Mixta de Procesamiento Mental (PPM) se diseñaron para reducir los efectos de la parcialidad de sexo y raza, y según la mayor parte de los informes, los autores de la prueba tuvieron éxito en estos

objetivos (Nolan, Watlington y Willson, 1989). Con base en los datos de estandarización, Kaufman, Kamphaus y Kaufman (1985) informaron de diferencias pequeñas (del orden de cinco puntos) entre la PPM de personas de raza blanca y de miembros de grupos minoritarios en la K-ABC. Ésta es una diferencia mucho más pequeña que la que se encuentra por lo común con pruebas como el WISC-III y la Stanford-Binet, donde son comunes las diferencias del orden de 15 puntos de CI a favor de las personas de raza blanca.

Valencia y Rankin (1988) sometieron a prueba a 76 niños de raza blanca y 90 méxico-estadounidenses de quinto y sexto grados e informaron que casi no había diferencia en la Puntuación Mixta de Procesamiento Mental (100 contra 98, respectivamente), aunque se encontró una gran diferencia en la Escala de Rendimiento (103 contra 91). Knight, Baker y Minder (1990) informan de una diferencia comparable entre la K-ABC y la SB:CE con 30 niños afroestadounidenses, estudiantes de primaria, que presentaban discapacidad para el aprendizaje (PPM de 83 contra Puntuación Mixta de 84 en la SB:CE).

En el cuadro 5–7 se describen las escalas y subpruebas de la K-ABC. Además de las Escalas Simultánea, Secuencial y de Rendimiento, se puede calcular una Escala No Verbal complementaria con seis subpruebas (de los grupos simultáneo y secuencial) que no requieren de palabras. Las subpruebas de la Escala No Verbal incluyen las siguientes:

Movimientos manuales.
Reconocimiento de rostros.
Triángulos.
Analogías con matrices.
Memoria espacial.
Series de fotografías.

En el caso de las subpruebas No Verbales, el examinador demuestra cada tarea mediante ejemplos o mímica. La Escala No Verbal es apropiada para examinar a niños bilingües o de inmigración reciente, cuyas habilidades en el idioma español podrían ser débiles. Además, los niños con deterioro auditivo, trastornos del habla o del lenguaje pueden evaluarse de manera imparcial con esta escala.

La K-ABC se estandarizó con una muestra nacional estratificada de 8 000 niños seleccionados de modo cuidadoso para representar los datos del censo de EUA para 1980 en cuanto a sexo, región geográfica, educación de los padres, tamaño de la comunidad y categoría étnica (raza blanca, afroestadounidenses, hispanos, otros). Un rasgo poco común y bienvenido de la muestra de estandarización es el énfasis en niños asignados a educación especial y a programas para inteligencia dotada y para niños talentosos (aproximadamente 7% de la muestra normativa). Por otra parte, las normas socioculturales complementarias para raza y educación de los padres se obtuvieron de los resultados de prueba de 469 niños afroestadounidenses y 119 de raza blanca.

La confiabilidad de la K-ABC es bastante buena en términos generales, aunque algunas subpruebas poseen coeficientes de consistencia interna apenas aceptables, en especial con los individuos más jóvenes. En el caso de niños preescolares, los valores medios van desde .72 para Ventana mágica hasta .88 para Recuerdo de números. En cuanto a los niños en edad escolar, los valores medios van de .71 para Cierre Gestáltico a .85 para Analogías con Matrices. Por otro lado, la confiabilidad de las puntuaciones de Escala y de la puntuación mixta es muy fuerte. Por ejemplo, la confiabilidad *test-retest* de la Escala de Rendimiento es de .93 para niños preescolares y de .97 para aquéllos en edad escolar. La Puntuación Mixta de Procesamiento Mental tiene una confiabilidad de .90 con niños preescolares y de .96 con aquéllos en edad escolar.

Los estudios sobre la validez de la K-ABC presentan una imagen heterogénea, con buen apoyo para la validez convergente de discriminación, fuerte confirmación de la validez relacionada con el criterio y buena sustentación para los cambios apropiados a la edad en las puntuaciones de prueba (pertinente para la validez de constructo). No obstante, los estudios analítico-factoriales sobre validez de constructo muestran resultados mixtos y conflictivos. Kamphaus,

Cuadro 5–7. Descripción de las subpruebas de la K-ABC

Escala de Procesamiento Secuencial

Movimientos manuales: el niño debe copiar la secuencia precisa de golpes en la mesa con el puño, palma o lado de la mano ejecutados por el examinador.

Recuerdo de números: muy similar a la prueba tradicional de retención de dígitos, excepto que al examinador se le instruye a no dejar caer su voz después de decir el último dígito.

Orden de palabras: mide la capacidad del niño para señalar a siluetas de objetos comunes en igual orden que los mismos objetos nombrados por el examinador.

Escala de Procesamiento Simultáneo

Ventana mágica: requiere que el niño identifique y nombre un objeto cuya ilustración se gira detrás de una abertura estrecha, de modo que sólo se expone una fracción de la ilustración en algún momento específico.

Reconocimiento de rostros: el niño debe prestar una estrecha atención a uno o dos rostros en una fotografía mostrada de manera breve y después seleccionar el rostro o rostros correctos en una fotografía de grupo.

Cierre gestáltico: el niño debe nombrar o describir de manera precisa una mancha parcialmente terminada, similar a una mancha de tinta. Esta subprueba mide la capacidad para llenar mentalmente los espacios vacíos para formar una *gestalt*.

Triángulos: el niño debe ensamblar varios triángulos idénticos de goma (amarillos en un lado, azules en el otro) para equipararlos con una ilustración de un diseño abstracto.

Analogías con matrices: con fichas de vinilo, el niño debe seleccionar la ilustración o diseño que mejor complete una matriz de 5 por 5 cm que expresa una analogía visual.

Memoria espacial: el niño debe recordar la localización de ilustraciones ordenadas de manera aleatoria en una página.

Series de fotografías: el niño debe ordenar en una secuencia temporal adecuada una disposición de fotografías dispuestas de manera aleatoria. Esta subprueba es similar a Ordenación de dibujos en las escalas Wechsler, excepto que la tarea debe resolverse sin manipulación física, lo cual elimina la tensión innecesaria sobre la realimentación visomotora.

Escala de Rendimiento

Vocabulario expresivo: el niño debe nombrar el objeto representado en una fotografía.

Rostros y lugares: el niño debe nombrar a una persona muy conocida, personaje ficticio o lugar representado en una fotografía.

Aritmética: una prueba de habilidades básicas de cálculo y de capacidades aritméticas relacionadas con la escuela.

Acertijos: el niño debe inferir el nombre de un concepto concreto o abstracto con base en una lista de sus características.

Lectura/Decodificación: una prueba de identificación de letras y de reconocimiento/pronunciación de palabras.

Lectura/Comprensión: el niño debe demostrar comprensión de lectura al seguir instrucciones dadas en oraciones.

Beres, Kaufman y Kaufman (1996) proporcionan una amplia revisión de, literalmente, docenas de estudios de validez con la K-ABC.

Las escalas Simultánea y Secuencial muestran el patrón esperado de correlaciones con los factores simultáneo y sucesivo en otras baterías de prueba (p. ej., Das y Mensink, 1989), lo cual indica buena validez convergente y de discriminación para estas escalas globales de la K-ABC. Cooley y Ayres (1985) encontraron correlaciones apropiadamente fuertes entre las puntuaciones de la K-ABC y medidas de rendimiento (validez convergente) y relaciones apropiadamente insignificantes entre las puntuaciones de la K-ABC y medidas de ansiedad en la infancia (validez de discriminación). También se encontró un descubrimiento un tanto singular: la Puntuación Mixta de Procesamiento Mental de la K-ABC se correlacionó en –.51 con la escala de Hiperactividad de la *Achenbach Child Behavior Checklist* (Lista de Verificación Achenbach de Conducta Infantil) (la correlación es negativa porque una alta puntuación en Hiperactividad indica conducta disfuncional). Esta lista de verificación bien validada consiste en reactivos que califican los padres (p. ej., el niño no puede concentrarse, no puede prestar atención durante largo tiempo, no es agradable para otros niños). En apariencia, la K-ABC detecta en cierto grado las capacidades de atención. Quizá la superposición es indirecta, donde la deficiencia en capacidades de atención conduce a la reducción en el logro de las habilidades intelectuales que se miden con la K-ABC.

La validez de esta prueba se sustenta también en los 43 estudios citados y resumidos en el Manual Interpretativo, incluyendo numerosos estudios de correlación con otras pruebas (Kaufman y Kaufman, 1983). Estas correlaciones varían de manera amplia, pero en general dan apoyo a la validez de la K-ABC, al menos en cuanto a la puntuación mixta de procesamiento mental. En este sentido, las conclusiones de Obrzut, Obrzut y Shaw (1984) y de Naglieri (1985) son típicas. Con muestras independientes de estudiantes con discapacidad para el aprendizaje y con retraso mental educable, estos dos estudios informaron correlaciones casi idénticas de .80 y .83, respectivamente, entre el CI del WISC-R y las puntuaciones mixtas de la K-ABC.

Como recordará el lector del capítulo sobre validez, una manera de demostrar la validez de constructo de una prueba consiste en mostrar que los cambios relacionados con la edad en las puntuaciones naturales son regulares, sensatos y consistentes con la teoría. Reynolds, Willson y Chatman (1984) correlacionaron la edad y las puntuaciones naturales de la muestra de estandarización (N = 2 000) y de una muestra adicional de personas afroestadounidenses y de raza blanca (N = 615). Todas las correlaciones entre la edad y las puntuaciones naturales fueron sumamente significativas. Y más importante aún, no ocurrieron diferencias significativas en la magnitud de estas relaciones en función de los grupos por raza o edad, lo cual sustenta la validez de constructo de la K-ABC como medida de aptitud y rendimiento asociada con el desarrollo.

Los estudios analítico-factoriales de las dimensiones secuencial y simultánea en las que se basa la K-ABC han producido resultados conflictivos (Kamphaus, 1990). En el lado positivo, los autores de la prueba citan conclusiones tomadas de la muestra de estandarización que parecen confirmar la distinción entre secuencial y simultáneo (Kaufman y Kaufman, 1983; Kaufman, Kaufman, Kamphaus y Naglieri, 1982). Algunos estudios independientes también han informado de resultados similares (p. ej., McCallum, Karnes y Oehler-Stinnett, 1985).

Los críticos reconocen que la K-ABC es una buena medida de la inteligencia general, pero se muestran dudosos acerca de la distinción entre procesamiento simultáneo y secuencial como base para comprender el desempeño en la prueba. Por ejemplo, Strommen (1988) emprendió un análisis factorial confirmatorio de la K-ABC con el propósito específico de examinar la hipótesis de que los factores que componen a la prueba sólo se correlacionan de manera moderada, lo cual constituye una de las afirmaciones clave de los autores de la prueba (Kaufman y Kaufman, 1983). Este investigador concluyó que los factores que subyacen a la K-ABC

están intercorrelacionados de manera sustancial en todos los niveles por edad, lo cual arroja dudas sobre la existencia independiente de los procesos secuencial y simultáneo en la prueba. Estos constructos modernos bien pueden resultar vino viejo en pellejos nuevos, nada más que una redenominación de la dicotomía familiar entre formas sumamente intercorrelacionadas de razonamiento verbal y no verbal. Para una interpretación más positiva de los estudios analítico-factoriales de la K-ABC, véase el artículo de reseña de Kamphaus y colaboradores (1996).

Una segunda crítica tiene que ver con la designación de las seis subpruebas como pruebas de rendimiento. Anastasi (1985) señala que una prueba puede considerarse apropiadamente como prueba de rendimiento sólo cuando se le vincula de manera estrecha con un contenido educativo específico. No obstante, los autores de la K-ABC hicieron esfuerzos especiales para separar las pruebas de rendimiento de cualquier contenido curricular específico. Estas subpruebas se parecen más a las medidas tradicionales de inteligencia que a las de rendimiento académico. ¡De hecho, con base en las cargas factoriales en el primer factor sin rotación, muchos investigadores han concluido que las subpruebas de Rendimiento proporcionan una mejor medida de la inteligencia general que aquéllas de procesamiento mental! (Kline, Guilmette, Snyder y Castellanos, 1992). Esta controversia se basa, en parte, en posturas filosóficas divergentes en cuanto a la naturaleza de la inteligencia y no es probable que se resuelva a través de la investigación (Reynolds, 1994b).

Otra preocupación es que la K-ABC no detecta de manera suficiente las habilidades verbales (Sattler, 1988). A pesar de estas controversias, la K-ABC ofrece un enfoque único e interesante de la evaluación de la inteligencia infantil y posee normas muy elevadas de calidad técnica. En tanto los examinadores incluyan las subpruebas de Rendimiento —que detectan la inteligencia general en grado sustancial— la K-ABC puede proporcionar un enfoque nuevo y valioso de la evaluación psicoeducativa.

ESCALAS MCCARTHY DE APTITUDES Y PSICOMOTRICIDAD PARA NIÑOS

Las Escalas McCarthy de Aptitudes y Psicomotricidad para Niños constituyen una prueba de inteligencia de aplicación individual, diseñada para niños entre los 2 años y medio y los 8 años y medio de edad (McCarthy, 1972). La prueba consiste en 18 subpruebas independientes, como se listan en el cuadro 5–8. Éstas contribuyen a las puntuaciones en cinco escalas, cada una se obtiene de 3 a 7 subpruebas: Verbal, Perceptual-Ejecución, Cuantitativa, Memoria y Motora. Además, se puede calcular un Índice Cognitivo General con media de 100 y *DE* de 16 a partir de 15 subpruebas. La prueba se diseñó para proporcionar un mejor entendimiento tanto de los niños normales como de aquellos que tienen discapacidad para el aprendizaje. McCarthy (1972) enfatizó las consideraciones funcionales, como el deseo de identificar las debilidades cognitivas clínica y educativamente relacionadas, como el principal criterio para la selección de reactivos y para los agrupamientos de las subpruebas en las Escalas McCarthy.

La muestra de estandarización de 1 032 niños consistió de aproximadamente 100 individuos en cada grupo de edad, con incrementos de medio año desde los 2 años y medio hasta los 5 años y medio de edad e incrementos de un año desde los 5 años y medio hasta los 8 años y medio. En cada nivel de edad, la muestra se estratificó de manera aproximada en las siguientes variables, de acuerdo con el censo de Población de EUA para 1970: sexo, raza (blanca-no blanca), región geográfica, nivel ocupacional del padre y residencia urbana-rural. Se excluyó a los niños con problemas mentales o emocionales graves y se incluyó a individuos bilingües sólo si podían hablar y comprender el Inglés. Por supuesto, un problema potencial con las escalas McCarthy es que los datos normativos, obtenidos a inicios del decenio de 1970, son bastante obsoletos.

Los datos de confiabilidad para las escalas McCarthy presentan una imagen heterogénea. El Índice Cognitivo General tiene un desempeño adecuado, con confiabilidades de división

Cuadro 5–8. Subpruebas y escalas de las Escalas McCarthy de Aptitudes y Psicomotricidad para Niños

Subpruebas	Escalas					
	Verbal	Ejecución perceptual	Cuantitativa	Memoria	Motora	Índice cognitivo general
Memoria pictórica	V			Mem		ICG
Conocimiento de palabras	V					ICG
Memoria verbal	V			Mem		ICG
Fluidez verbal	V					ICG
Analogías opuestas	V					ICG
Construcción con cubos		P				ICG
Solución de rompecabezas		P				ICG
Secuencia rítmica		P		Mem		ICG
Orientación derecha-izquierda		P				ICG
Dibujo de un diseño		P			Mot	ICG
Dibujo de un niño		P			Mot	ICG
Agrupamiento conceptual		P				ICG
Preguntas con números			C			ICG
Memoria numérica			C	Mem		ICG
Conteo y clasificación			C			ICG
Coordinación de piernas					Mot	
Coordinación de brazos					Mot	
Acción imitativa					Mot	

en mitades que promedian cerca de .93 y coeficientes *test-retest* a un mes que tienen un promedio aproximado de .90. Las confiabilidades de división en mitades para las cinco escalas van de .79 a .88, mientras los coeficientes *test-retest* van de .69 a .89. Las confiabilidades para las 18 subpruebas individuales son sustancialmente menores, de modo que se advierte a los examinadores que no coloquen demasiado énfasis en los patrones y diferencias de subprueba.

Por desgracia, no se ha confirmado la derivación con base clínica de las cinco escalas McCarthy a través de estudios analítico-factoriales, lo cual deja en duda la validez de este instrumento. Aunque los cinco factores (que corresponden a las cinco escalas) se encontraron en todos los grupos por edad dentro de la muestra de estandarización (Kaufman, 1975), estudios posteriores no han reproducido estas conclusiones originales. Por ejemplo, Forns-Santacana y Gómez-Benito (1990) encontraron cinco factores en una muestra de 141 niños de 4 y 5 años de edad, pero éstos no corresponden a la clasificación propuesta por McCarthy. Otros investigadores informan de casos similares en los que no se pudo corroborar la distribución original de las subpruebas. Por ejemplo, Keith y Bolen (1980) encontraron sólo tres factores

en una muestra de 300 niños con edades de 6 a 8 años y medio de edad: cognitivo general, verbal y motor.

La confusión acerca de la estructura factorial de las escalas McCarthy indica que los examinadores deben ser precavidos en el análisis de un perfil que se base en las cinco escalas antes mencionadas (Verbal, Perceptual-Ejecución, Cuantitativa, Memoria y Motora). En muchas muestras y para algunos grupos por edad, las escalas pueden ser mejores como medidas de la capacidad cognitiva general que de las capacidades específicas designadas por los nombres de las escalas (Sattler, 1988).

En el lado positivo, las escalas McCarthy funcionan muy bien como medios de predicción de la preparación para la escuela y del rendimiento escolar posterior en el caso de niños en el jardín de infantes. Massoth y Levenson (1982) sometieron a prueba a 33 niños con la escala McCarthy en el periodo de otoño del jardín de infantes y correlacionaron estas puntuaciones con los resultados de la prueba de preparación para la lectura aplicada un año después y también con los niveles de rendimiento a la terminación del primer grado de primaria. Curiosamente, las correlaciones más fuertes se obtuvieron con la Escala Cuantitativa, mientras la Escala Verbal tuvo un resultado deficiente como medio de predicción de la preparación para la escuela o del rendimiento en lectura (cuadro 5–9). Las capacidades perceptuales y analíticas que miden las escalas McCarthy parecen ser mejores medios de predicción de la preparación y del rendimiento en lectura que las tareas verbales. Un seguimiento a los seis años con los mismos individuos reveló correlaciones sorprendentemente altas entre las escalas McCarthy aplicadas en el jardín de infantes y el rendimiento escolar en el sexto grado (Massoth, 1985). La Escala Cuantitativa mostró la correlación más fuerte con las calificaciones del curso ($r = .60$), mientras que la Escala Verbal constituyó un medio de predicción débil ($r = .40$). La Escala Cuantitativa parecería ser una excelente prueba de detección para niños preescolares.

En resumen, las escalas McCarthy proporcionan un índice valioso y adecuado para la predicción del funcionamiento intelectual, en especial con niños en el rango de 5 a 6 años de edad. El instrumento también es una excelente herramienta para la evaluación de la inteligencia general, aunque puede subestimar el funcionamiento de preescolares, niños con discapacidad para el aprendizaje y con retraso mental. En la actualidad, las normas para la prueba son sumamente obsoletas. A pesar de esto, las puntuaciones generales en la escala McCarthy corresponden de manera estrecha a los CI Totales del WPPSI-R (Karr, Carvajal, Elser y Bays, 1993). Aun así, las escalas McCarthy necesitan revisión y reestandarización.

Cuadro 5–9. Correlaciones entre las puntuaciones de la Escala McCarthy y la preparación, y el rendimiento en lectura de 33 niños en el jardín de infantes

Escala McCarthy	Preparación para la Lectura de Macmillan	Prueba Metropolitan de Rendimiento
Verbal	.33	.16
Ejecución-Perceptual	.39*	.37*
Cuantitativa	.64**	.50**
General Cognitiva	.53**	.39*
Memoria	.39*	.28
Motora	.31	.35*

*$p < .05$
**$p < .01$
Fuente: Adaptado con autorización de Massoth, N. A. y Levenson, R. L. (1982). The McCarthy Scales of Children's Abilities as a predictor of reading readiness and reading achievement. *Psychology in the Schools, 19,* 239-296.

ESCALAS DE CAPACIDAD DIFERENCIAL

Las *Differential Ability Scales* (DAS) son una adición reciente a la evaluación individual de la inteligencia que es digna de un breve comentario (Elliott, 1990, 1997). La DAS cubre un rango de edades de los 2 años y medio a los 18 años en tres baterías superpuestas: preescolar

inferior (edades de 2:6 a 3:5), preescolar superior (edades de 3:6 a 5:11) y edad escolar (edades de 6:0 a 17:11). Aquí se presenta la batería preescolar superior.

Las subpruebas de la batería preescolar incluyen subpruebas "esenciales" y "diagnósticas". Las primeras están altamente saturadas del factor *g* y se utilizan para obtener dos puntuaciones de área (Verbal y No Verbal) y una puntuación mixta global conocida como Capacidad Conceptual General (CCG). Las puntuaciones de área y la CCG se basan en una media de 100 y desviación estándar de 15. Las subpruebas diagnósticas miden memoria a corto plazo y velocidad de procesamiento de información. Se les utiliza sólo para el análisis clínico. Las subpruebas diagnósticas dependen menos del factor *g* y, por tanto, no figuran en la puntuación mixta global. Las subpruebas de la DAS se describen en el cuadro 5–10.

La confiabilidad de las puntuaciones de la DAS es encomiable para un instrumento que se utiliza en el nivel preescolar. En el caso de preescolares, se informa que la confiabilidad de la CCG es de .90 a .94. Para preescolares mayores (3 años y medio a 6 años de edad) las puntuaciones agrupadas Verbales y No Verbales muestran confiabilidades de .88 y .89,

Cuadro 5–10. Subpruebas de la Batería Preescolar DAS

Subprueba	Capacidades Medidas	Contribución a la puntuación mixta
Subpruebas esenciales		
Comprensión verbal	Lenguaje receptivo, comprensión de instrucciones verbales	Verbal, CCG
Vocabulario de nombres	Lenguaje expresivo, conocimiento de nombres	Verbal, CCG
Semejanzas con dibujos	Razonamiento no verbal, equiparación de ilustraciones con temas comunes	No verbal, CCG
Construcción de patrones	Visualización no verbal, espacial, con cubos y cuadrados de colores	No verbal, CCG
Copiado	Copiado de diseños, coordinación motora fina; equiparación visoespacial	No verbal, CCG
Primeros conceptos de número	Conocimiento de conceptos de número y cuantitativos	CCG
Subpruebas diagnósticas		
Construcción con cubos	Orientación espacial, equiparación visoperceptual con cubos	n/a
Equiparación de formas parecidas a letras	Relaciones espaciales, discriminación visual de formas similares	n/a
Recuerdo de dígitos	Memoria auditiva a corto plazo de secuencias numéricas	n/a
Recuerdo de objetos	Aprendizaje a corto plazo y recuerdo verbal de ilustraciones	n/a
Reconocimiento de dibujos	Memoria visual a corto plazo, reconocimiento de objetos familiares	n/a

n/a = no aplica.

Cuadro 5–11. Correlaciones entre las puntuaciones mixtas de la DAS y el WISC-III de 53 niños con discapacidad para el aprendizaje

Puntuación mixta de la DAS	Puntuación mixta del WISC-III			DAS	
	CIV	CIE	CIT	Media	DE
Verbal	.77	.52	.72	90.2	12.0
Razonamiento no verbal	.55	.65	.67	83.5	12.5
Espacial	.50	.67	.64	93.6	17.0
CCG	.68	.71	.78	87.2	14.8
Media del WISC-III	89.4	93.2	89.7		
DE del WISC-III	13.8	14.2	13.2		

Fuente: Reimpreso con autorización de Dumont, R., Cruse, C., Price, L. y Whelley, P. (1996). The relationship between the Differential Ability Scales (DAS) and the Wechsler Intelligence Scale for Children–Third Edition (WISC-III) for students with learning disabilities. *Psychology in the Schools, 33,* 203-209.

respectivamente. Los estudios sobre validez concurrente prestan un fuerte apoyo a la DAS, con correlaciones en los rangos de .70 y .80 con otras medidas preescolares de la inteligencia y rendimiento (Elliott, 1990ab). Un estudio de Dumont, Cruse, Price y Whelley (1996) provee una muy fuerte sustentación para la validez de la DAS al proporcionar un patrón confirmatorio de las correlaciones entre este instrumento y el WISC-III con 53 niños en los que se había identificado discapacidad para el aprendizaje. Los resultados se resumen en el cuadro 5–11 y muestran que los componentes similares se correlacionan en mayor medida que aquéllos diferentes en las dos pruebas. También, dicho cuadro revela que las puntuaciones totales son muy similares, en promedio, entre la DAS (CCG media de 87.2) y el WISC-III (CI medio de 89.7). Elliott (1997) describe estudios adicionales de validez para este magnífico instrumento.

UTILIDAD PRÁCTICA DE LA EVALUACIÓN DE LACTANTES Y PREESCOLARES

En general, la historia de la evaluación infantil ha mostrado una y otra vez que las puntuaciones de prueba que se obtienen en los primeros uno o dos años de vida muestran una validez predictiva mínima. Por ejemplo, en su reseña acerca de las pruebas de inteligencia con lactantes, Goodman (1990) concluyó:

> Si el pronóstico exitoso de la inteligencia de adolescentes y adultos a partir de las puntuaciones en la infancia temprana es uno de los principales logros de la psicología aplicada, entonces la incapacidad para predecir la inteligencia de la primera infancia a partir de la lactancia puede considerarse como uno de sus mayores fracasos.

Dado el pésimo historial de fracasos repetidos en cuanto a validez predictiva, debería hacerse una difícil pregunta: ¿Cuál es el propósito y utilidad práctica de la evaluación de lactantes? De hecho, las pruebas con éstos tienen una función importante, pero limitada, que representar. Se regresará a esta cuestión después de revisar los estudios predictivos.

VALIDEZ PREDICTIVA DE LAS PRUEBAS CON LACTANTES Y PREESCOLARES

Cuando se trata de muestras heterogéneas de niños normales, el resultado general es que las puntuaciones de pruebas con lactantes se correlacionan en sentido positivo, pero de modo poco impresionante, con las puntuaciones de prueba en la infancia (Goodman, 1990; McCall, 1979). Unos cuantos estudios muestran un tono más optimista (p. ej., Wilson, 1983), pero la

mayoría de los investigadores concuerda con la conclusión de McCall (1976):

> En términos generales, esencialmente no existe correlación entre el desempeño durante los primeros seis meses de vida y la puntuación de CI después de los cinco años de edad; las correlaciones se colocan de manera predominante en el nivel de 0.20 para las evaluaciones realizadas entre los 7 y 18 meses de vida cuando se trata de predecir el CI en el periodo de 5 a 18 años de edad, y no es sino hasta los 19 a 30 meses que la prueba con lactantes pronostica el CI posterior en el rango de 0.40 a 0.55.

McCall (1979) volvió a confirmar su conclusión original en una reseña posterior que se ha resumido aquí. El lector observará en el cuadro 5–12 que las correlaciones entre las puntuaciones de pruebas con lactantes y con niños en edad escolar no superan al .40 hasta que los individuos tienen, cuando menos, 19 meses de edad en la prueba inicial.

Las conclusiones con pruebas para preescolares tienen un carácter un tanto más positivo. La correlación entre los resultados de pruebas preescolares y el CI posterior es, por lo común, fuerte, significativa y representativa. La manera más sencilla de investigar esta cuestión consiste en medir la estabilidad de los resultados de CI en estudios longitudinales. En el cuadro 5–13 se resume la estabilidad de las puntuaciones de CI de niños en la Stanford-Binet de una edad a otra. Estos datos provienen del Estudio Longitudinal Fels, una de las primeras investigaciones de seguimiento del desarrollo intelectual y emocional de niños (Sontag, Baker y Nelson, 1958) que se considera clásica. La correlación más baja en este cuadro es .43 y se encuentra entre el CI evaluado a los cuatro años y de nuevo a los 12 años

Cuadro 5–12. Resumen de correlaciones entre puntuaciones de prueba de inteligencia para lactantes y niños en individuos normales

	Edad de prueba en la infancia (años)		
Edad de la prueba inicial en la lactancia (meses)	3–4	5–7	8–18
1–6	.21	.09	.06
7–12	.32	.20	.25
13–18	.50	.34	.32
19–30	.59	.39	.49

Fuente: Adaptado con autorización de McCall, R. B. (1979). The development of intellectual functioning in infancy and the prediction of later IQ. En J. D. Osofsky (Ed.), *Handbook of infant development*. Nueva York: John Wiley.

Cuadro 5–13. Estabilidad del CI de los 3 a los 12 años de edad

	Edad de la repetición de la prueba								
Edad de la prueba inicial	4	5	6	7	8	9	10	11	12
3	.83	.72	.73	.64	.60	.63	.54	.51	.46
4		.80	.85	.70	.63	.66	.55	.50	.43
5			.87	.83	.79	.80	.70	.63	.62
6				.83	.79	.81	.72	.67	.67
7					.91	.83	.82	.76	.73
8						.92	.90	.84	.83
9							.90	.82	.81
10								.90	.88
11									.90

Fuente: Adaptado con autorización de Sontag, L.W., Baker, C. y Nelson, V. (1958). Mental growth and personality development: A longitudinal study. *Monographs of the Society for Research in Child Development, 23*, (No. 68 completo). Copyright © de The Society for Research in Child Development, Inc.

de edad. Lo que se destaca del cuadro es la fortaleza del vínculo entre el CI en la edad preescolar y en la infancia posterior. Entre mayor sea el niño cuando se le somete a prueba por primera vez, más fuerte será la relación con el CI posterior. De hecho, los resultados sugieren que, en promedio, el CI se vuelve razonablemente estable hacia los ocho años de edad.

De manera colectiva, estas conclusiones confirman que, en general, las pruebas con lactantes tienen un valor pronóstico deficiente, mientras que aquellas con preescolares tienen un desempeño moderado en la predicción de la inteligencia posterior. Esto conduce a la pregunta planteada al inicio de esta sección: ¿Cuál es el propósito y utilidad práctica de la evaluación de lactantes?

UTILIDAD PRÁCTICA DE LAS ESCALAS BAYLEY–II Y DE OTRAS ESCALAS PARA LACTANTES

El empleo más importante y justificable de las pruebas con lactantes es la detección de discapacidades del desarrollo. Aunque las pruebas existentes para la etapa de la lactancia son deficientes como medios de predicción de la inteligencia en la infancia, una excepción a esta regla se encuentra en el caso de lactantes que obtienen una puntuación muy baja en pruebas como la Bayley-II y otros instrumentos de detección. Por ejemplo, los lactantes que tienen puntuaciones a dos desviaciones estándar por debajo de la media en la escala Bayley, en particular en la Escala Mental, tienen una elevada probabilidad de obtener después en su vida resultados de prueba en los rangos de retraso mental (Self y Horowitz, 1979; Goodman, Malizia, Durieux-Smith, MacMurray y Bernard, 1990).

Con niños en situación de riesgo, la correlación entre las puntuaciones de prueba en la lactancia y el CI posterior en la niñez es mucho más fuerte que para las muestras de niños normales. McCall (1983) determinó que la correlación mediana entre las puntuaciones de prueba en la lactancia y el CI en la infancia en un seguimiento a siete años era un robusto .48. La conclusión más consistente es que una puntuación muy baja en una prueba para lactantes —dos desviaciones estándar por debajo de la media o más baja— predice de manera precisa un CI bajo en la infancia (Frankenburg, 1985). Por ejemplo, los estudios con la *Denver Developmental Screening Test-Revised* (Prueba Denver de Detección del Desarrollo–Revisada; desde entonces publicada como Denver–II) reveló una tasa falsa positiva de sólo 5 a 11%, lo cual significa que los lactantes y preescolares identificados como niños en riesgo pocas veces lograban un funcionamiento en el rango normal. Los estudios con las escalas Bayley también se conforman a este patrón (p. ej., VanderVeer y Schweid, 1974).

NUEVOS ENFOQUES DE LA EVALUACIÓN CON LACTANTES

Lewis ha afirmado que las pruebas tradicionales para lactantes pasan por alto las conductas iniciales de procesamiento de información, como la memoria de reconocimiento y la atención hacia el ambiente, que podrían predecir mejor la función cognitiva en la infancia (Lewis y Sullivan, 1985). En un estudio, la simple habituación visual a un estímulo novedoso (medida por la duración de la fijación), evaluada a los tres meses de edad, se correlacionó en .61 con la puntuación obtenida en la Escala Mental de las escalas Bayley a los 24 meses de edad (Lewis y Brooks-Gunn, 1981). Con un paradigma similar, Fagan ha informado de resultados comparables (Fagan, 1984; Fagan y Shepherd, 1986). Por ejemplo, en un estudio, este investigador sometió a prueba con el método de habituación la memoria de reconocimiento de lactantes a las edades de 4 a 7 meses (Fagan y McGrath, 1981). En este estudio, los lactantes observaron primero una fotografía del rostro de un bebé durante un periodo corto y después se les mostró la misma fotografía junto con una desconocida (p. ej., la imagen de un hombre calvo). Los investigadores llevaron un registro cuidadoso de cuál era la fotografía que los lactantes miraban más; la lógica detrás del

procedimiento es sencilla: Mirar principalmente la nueva fotografía significa que el lactante reconoce la anterior; es decir, un lactante con buena memoria de reconocimiento prefiere mirar algo nuevo. Por tanto, la preferencia por la novedad —como se mide a través del tiempo de fijación visual en la nueva imagen— se vuelve un índice de la memoria inicial de reconocimiento. Años después, los investigadores aplicaron la *Peabody Picture Vocabulary Test* (Prueba Peabody de Vocabulario con Figura; PPVT) para estimar la inteligencia en la infancia temprana. Las puntuaciones de memoria de reconocimiento en la lactancia y las puntuaciones en la PPVT en la infancia temprana se correlacionaron en .37 a los cuatro años de edad y en .57 a los siete. Es probable que estas correlaciones subestimen la validez predictiva de las pruebas de memoria para lactantes, en el sentido de que el índice de memoria en la lactancia fue un procedimiento poco confiable basado en un pequeño número de reactivos de prueba. Lo que es más, los investigadores evaluaron a lactantes normales, lo cual diminuyó las correlaciones entre el medio de predicción y el criterio.

Las medidas cognitivas para lactantes representan una gran promesa como medios de predicción de la inteligencia en la infancia (Bornstein, 1994; Fagan y Haiken-Vasen, 1997). En los siguientes años se podría ser testigo del surgimiento de tipos completamente nuevos de instrumentos de evaluación para lactantes que se basen en la medición de la memoria temprana, habituación y capacidades de atención, en lugar de las capacidades sensoriomotoras. Un primer paso en esta dirección es la *Fagan's Test of Infant Intelligence* (Prueba Fagan de Inteligencia para Lactantes; Fagan y Shepherd, 1986), que es un instrumento sencillo basado en los métodos antes descritos para la medición de la preferencia por la novedad y la memoria de reconocimiento de los lactantes. Esta prueba produce una puntuación mixta que se basa en la preferencia por la novedad —como se mide a través del tiempo de fijación visual en una nueva imagen— promediada en varios ensayos. El procedimiento muestra una concordancia intercalificadores muy alta (O'Neill, Jacobson y Jacobson, 1994).

Los estudios iniciales sobre la validez de la *Fagan's Test of Infant Intelligence* (FTII) como medio de predicción de la inteligencia en la infancia han obtenido resultados mixtos. En una muestra de 200 lactantes, las puntuaciones de la FTII obtenidas entre los 7 y 9 meses de edad se correlacionaron sólo en .32 con el CI de la Stanford-Binet a la edad de tres años (DiLalla, Thompson, Plomin y colaboradores, 1990). En otro estudio reciente, las correlaciones globales entre las puntuaciones de la FTII, obtenidas de los 7 a los 9 meses de edad, y el CI del WPPSI-R a los cinco años de edad, se colocaron alrededor del .2 en dos muestras noruegas de niños sanos (Andersson, 1996). Estas correlaciones no prestan apoyo al empleo de la prueba como herramienta de detección en poblaciones que no se consideren en riesgo. Sin embargo, la prueba puede funcionar mejor si se le utiliza con lactantes en riesgo. De todas maneras, se requiere mayor investigación antes de abandonar las medidas tradicionales para lactantes en favor de la prueba Fagan y otros métodos similares.

RESUMEN

1. El periodo de la lactancia y de la edad preescolar se extiende desde el nacimiento hasta aproximadamente los seis años. Una aplicación importante de las pruebas con lactantes y preescolares consiste en ayudar a responder preguntas sobre demoras del desarrollo. La mayor parte de las pruebas con lactantes (desde el nacimiento hasta los dos años y medio) tienen cargas notables en habilidades sensoriales y motoras. Las pruebas para preescolares (2 años y medio a 6 años de edad) tienden a dirigirse en

grado significativo hacia las habilidades cognitivas.

2. Los Inventarios Gesell del Desarrollo (GDS) estiman el progreso relacionado con el desarrollo de lactantes desde las cuatro semanas hasta los 60 meses de edad.
3. Las *Ordinal Scales of Psychological Development* se diseñaron como una medida con fundamentación piagetiana del desarrollo intelectual (desde las 2 semanas hasta los 2 meses de edad). Las escalas miden el desarrollo de la permanencia de objeto, medios-fines, imitación vocal y de ademanes, causalidad operacional, relaciones objetales en el espacio y esquemas para relacionarse con los objetos.
4. Las *Bayley Scales of Infant Development-II* evalúan el desarrollo mental y motor de niños de 1 a 42 meses de edad; tienen una estandarización cuidadosa y son sumamente confiables. Como en otras pruebas para lactantes, las puntuaciones muy bajas predicen un resultado de discapacidad intelectual en la infancia posterior, mientras las puntuaciones cercanas a la norma y superiores tienen poca validez predictiva.
5. La Escala Wechsler de Inteligencia para los Niveles Preescolar y Primaria–Revisada (WPPSI-R) se diseñó para niños de 3 a 7 años, 3 meses de edad. El WPPSI-R contiene tres subpruebas que no se encuentran en otras Escalas Wechsler: Frases (memoria oral), Diseños geométricos (copia de diseños) y Casa de los animales (asignación codificada de clavijas).
6. La Escala de Inteligencia Stanford-Binet: Cuarta Edición (SB:CE) es un instrumento útil para la evaluación preescolar. Aunque la prueba se diseñó para producir puntuaciones en cuatro factores, la solución de dos factores de Sattler (1988) para el informe de las puntuaciones de la SB:CE (Comprensión Verbal y Razonamiento No Verbal/Visualización) es el enfoque preferido con preescolares.
7. La Batería de Evaluación Kaufman para Niños (K-ABC), que se emplea con niños desde los 2 años y medio hasta los 12 años, 5 meses de edad, es una medida combinada de inteligencia y rendimiento que se basa en la distinción entre procesamiento secuencial (ordenación en serie o temporal de los estímulos) y procesamiento simultáneo (síntesis y organización de los estímulos en forma inmediata o global).
8. Las Escalas McCarthy de Aptitudes y Psicomotricidad para Niños se diseñaron para niños desde los 2 años y medio hasta los 8 años y medio. Las 18 subpruebas producen cinco subpuntuaciones diferentes y un Índice Cognitivo General (ICG) similar a un CI. Las subpuntuaciones incluyen las escalas verbal, perceptual-ejecución, cuantitativa, memoria y motora. Por desgracia, estas cinco áreas no se han confirmado a través de análisis factoriales independientes.
9. Diseñadas para niños de 2 años y medio a 17 años, 11 meses de edad, las *Differential Ability Scales* constan de 17 subpruebas cognitivas y tres pruebas de rendimiento, conormadas, que se dirigen a niños en edad escolar. La investigación inicial indica que la DAS produce puntuaciones confiables y razonablemente independientes de subprueba útiles para la evaluación de la discapacidad para el aprendizaje.
10. En general, las puntuaciones de pruebas para lactantes se correlacionan en sentido positivo, pero débilmente, con aquellas de pruebas para niños. Las puntuaciones de pruebas para lactantes deben interpretarse con precaución. Una excepción la constituyen las puntuaciones muy bajas en instrumentos como la Bayley-II, la cual predice de manera confiable la incapacidad del desarrollo en la infancia.
11. Las pruebas de memoria de reconocimiento en la lactancia parecen prometedoras como medios de predicción de la inteligencia en la infancia. Por ejemplo, en los estudios de Fagan, los índices de la habituación visual simple en la lactancia se correlacionan en .57 con las puntuaciones de vocabulario con figura a los siete años de edad. ■

Pruebas de inteligencia II: Pruebas individuales y grupales

CAPÍTULO 6

TEMA 6A
Pruebas individuales de inteligencia

Orientación hacia las pruebas individuales de inteligencia
Escalas Wechsler de inteligencia
Subpruebas Wechsler: descripción y análisis
Escala Wechsler de inteligencia para adultos-III
Escala Wechsler de inteligencia para el nivel escolar-III
Escala de inteligencia Stanford-Binet: Cuarta Edición
Pruebas Detroit de aptitud para el aprendizaje-4
Test breve de inteligencia de Kaufman (K-BIT)
Resumen
Términos y conceptos clave

El examen de la inteligencia es uno de los principales logros de la psicología de este siglo. En respuesta al éxito de las escalas Binet-Simon a principios del siglo XX, los psicólogos desarrollaron y depuraron docenas de pruebas individuales de inteligencia diseñadas con base en este instrumento innovador. También se observó un gran crecimiento en las pruebas grupales de inteligencia, fomentadas por la entusiasta aceptación de las pruebas Army Alfa y Beta durante y después de la primera Guerra Mundial. Sólo con algunas excepciones, las pruebas actuales de inteligencia, tanto individuales como grupales, son herederas de Binet, Simon y del programa de pruebas del ejército.

El propósito de este capítulo consiste en proporcionar una revisión de los enfoques más importantes en cuanto a pruebas individuales y grupales de inteligencia. En el Tema 6A se revisan las pruebas individuales sobresalientes y se termina el capítulo con una reseña de las pruebas grupales de inteligencia en el Tema 6B.

Aunque esta obra dedica tres capítulos completos al fascinante y emocional tema del examen de la inteligencia, no se afirma que la cobertura sea exhaustiva. Un análisis completo de las pruebas de inteligencia se encuentra simplemente más allá del campo de acción de ésta o cualquier otra referencia básica. Casi cada mes aparecen pruebas nuevas y revisadas, y cada año se publican miles de novedosos descubrimientos de investigación. Aquí se reseñan las pruebas más utilizadas o que ilustran avances interesantes en la teoría o el método. Los lectores pueden encontrar información acerca de pruebas adicionales en la serie del *Mental Measurements Yearbook*, que ahora publica cada 3 o 4 años el *Buros Institute* (p. ej., Mitchell, 1985; Conoley e Impara, 1995; Conoley y Kramer, 1989, 1992). La *Encyclopedia of Human Intelligence* (Sternberg, 1994) también es una buena fuente de información acerca de las pruebas individuales y grupales de inteligencia.

ORIENTACIÓN HACIA LAS PRUEBAS INDIVIDUALES DE INTELIGENCIA

Las pruebas individuales de inteligencia revisadas en este tema incluyen las siguientes:

- Escala Wechsler de inteligencia para adultos-III (WAIS-R)
- Escala Wechsler de inteligencia para el nivel escolar-III (WISC-III)
- Prueba de inteligencia Stanford-Binet: Cuarta Edición (SB:CE)
- *Detroit Test of Learning Aptitude*-4 (DTLA-4; Prueba Detroit de Aptitud para el Aprendizaje)
- Test Breve de Inteligencia de Kaufman (K-BIT)

Otro instrumento prometedor que no se revisa es *Kaufman Adolescent and Adult Intelligence Test* (KAIT; Prueba Kaufman de Inteligencia para Adolescentes y Adultos). Publicada en 1992, la KAIT ha llegado recientemente al escenario de las pruebas (Dumont y Hagberg, 1994; Shaugnessy y More, 1994). Kaufman y Kaufman (1997) enumeran varias ventajas de la KAIT, incluyendo su fundamentación psicométrica en la distinción G_c-G_f propuesta por John Horn y sus seguidores. La KAIT también es atractiva por su brevedad: la prueba proporciona índices sumamente confiables de inteligencia en dos terceras partes del tiempo que se requiere para la mayor parte de las baterías. Junto con las pruebas preescolares presentadas en tema anterior, es probable que los instrumentos antes listados expliquen 98 % de las evaluaciones intelectuales realizadas en EUA.

Las escalas Wechsler han dominado las pruebas de inteligencia en años recientes, pero de ninguna manera son las únicas opciones viables de evaluación individual. Muchos otros instrumentos miden igualmente bien la inteligencia general —incluso algunos dirían que mejor. Considérense las implicaciones de la siguiente observación, ahora familiar, de que en muestras grandes y heterogéneas, las puntuaciones en cualesquiera dos instrumentos establecidos (p. ej., escalas Wechsler, Stanford-Binet, McCarthy, Kaufman, etc.) se correlacionan por lo común de 0.80 a 0.90. Con frecuencia, la correlación entre dos instrumentos establecidos es casi tan alta como la correlación *test-retest* para cualquier instrumento por sí solo. Cuando el propósito es obtener una puntuación global, parecería que cualquier prueba de inteligencia establecida y con normas adecuadas sería suficiente.

Pero obtener una puntuación total no es la única meta de la evaluación. Además, por lo general, el examinador desea comprender el funcionamiento intelectual del individuo. Para este propósito, el CI total es importante, pero existen casos donde la puntuación global puede ser irrelevante o incluso engañosa. Para comprender el funcionamiento intelectual de alguien, el examinador también debería revisar las puntuaciones de subprueba en búsqueda de hipótesis que podrían explicar el funcionamiento único de ese individuo. Por supuesto, los examinadores necesitan emprender de manera cauta el análisis de subpruebas, con hallazgos basados en la investigación acerca de la naturaleza y

significado de la dispersión de las subpruebas en el caso de la prueba utilizada (Gregory, 1994b; McLean, Kaufman y Reynolds, 1989; McDermott, Fantuzo y Glutting, 1990).

Si el objetivo del examinador consiste en comprender el funcionamiento intelectual y no sólo determinar una puntuación total, las diferencias entre pruebas se vuelven bastante reales. Todo instrumento aborda la medición de la inteligencia desde una perspectiva diferente y produce un conjunto distintivo de puntuaciones de subprueba. Además, una prueba adecuada para un caso específico, podría tener un desempeño totalmente inadecuado en otro contexto. Por ejemplo, el WAIS-R tiene un desempeño admirable en el examen del retraso mental leve, pero contiene muy pocos reactivos simples para la evaluación de personas con discapacidades del desarrollo de moderadas a graves.

Un axioma central de la evaluación es que la elección de un instrumento de prueba debe basarse en el conocimiento de sus fortalezas y debilidades, según conciernen a la pregunta de referencia. Dicho de manera sencilla, ¡el examinador hábil no depende a ciegas de una sola prueba para todos los casos que se le refieren! En vez de ello, elige de manera flexible uno o más instrumentos, según las necesidades de evaluación que percibe en el individuo a examinar. Cada una de las pruebas que se analizan en este tema tienen sus méritos especiales y también sus deficiencias particulares. El usuario debe conocer estas facetas a fin de elegir el instrumento más adecuado para cada referencia única.

ESCALAS WECHSLER DE INTELIGENCIA

En el decenio de 1930, David Wechsler, un psicólogo del *Bellevue Hospital* de la ciudad de Nueva York, concibió una serie de instrumentos de gran sencillez que finalmente definieron las pruebas de inteligencia desde mediados hasta finales del siglo XX. Su influencia sobre el examen de la inteligencia se ve superada sólo por las contribuciones innovadoras de Binet y Simon. Es adecuado comenzar el estudio de las pruebas individuales con un resumen de la tradición de Wechsler, seguido de un análisis de los instrumentos en sí.

ORÍGENES DE LAS PRUEBAS WECHSLER

Wechsler comenzó a trabajar en su primera prueba en 1932, buscando diseñar un instrumento adecuado para examinar a los diversos pacientes que se referían a la sección psiquiátrica del *Bellevue Hospital* en Nueva York (Wechsler, 1932). Posteriormente, al describir el desarrollo de su primera prueba, escribió que "nuestra meta no era producir un conjunto de pruebas totalmente novedosas, sino seleccionar, de cualquier fuente disponible, una combinación de ellas que satisfaciera los requisitos de una escala eficaz para adultos" (Wechsler, 1939). De hecho, el contenido de sus escalas se inspiró en gran medida en las pruebas Army Alfa y Beta (Frank, 1983). Quienes han leído de manera detenida el libro *Psychological Examining in the United States*, editado por Yerkes (1921) justo después de la primera Guerra Mundial, podrían sorprenderse de descubrir que Wechsler se apropió de docenas de reactivos de prueba de esta fuente, muchos de los cuales han sobrevivido hasta la actualidad en las revisiones contemporáneas de las pruebas Wechsler. Este autor no era tanto un genio creativo como un individuo práctico, que diseñó un instrumento nuevo y útil a partir de elementos sobrantes de los intentos anteriores y descontinuados para examinar la inteligencia.

La primera de las pruebas Wechsler, llamada Escalas de inteligencia Wechsler-Bellevue, se publicó en 1939. Al analizar la fundamentación para su nueva prueba, Wechsler (1941) explicó que los instrumentos existentes, como la Stanford-Binet, eran deplorablemente inadecuados para la evaluación de la inteligencia adulta. La Wechsler-Bellevue se diseñó para rectificar varios defectos que se habían observado en pruebas anteriores:

- Los reactivos de prueba no tenían atractivo para las personas adultas.

- Demasiadas preguntas enfatizaban la mera manipulación de las palabras.
- Las instrucciones enfatizaban la velocidad a expensas de la precisión.
- Depender de la edad mental era irrelevante en el caso de pruebas con adultos.

Para corregir estas deficiencias, Wechsler diseñó su prueba específicamente para adultos, añadió reactivos de ejecución para equilibrar las preguntas verbales, redujo el énfasis en las preguntas con límite de tiempo e inventó un nuevo método para obtener el CI. En términos específicos, sustituyó la fórmula común:

$$CI = \frac{\text{edad mental}}{\text{edad cronológica}}$$

por una fórmula nueva relacionada con la edad

$$CI = \frac{\text{puntuación obtenida o real}}{\text{puntuación media esperada para esa edad}}$$

Esta nueva fórmula se basó en la interesante presuposición —expresada en forma de axioma— de que el CI permanece constante a medida que se envejece de manera normal, aunque la capacidad intelectual natural pueda cambiar o incluso disminuir. La suposición de la **constancia de CI** es básica para las escalas Wechsler. Como lo expresó el mismo autor (1941):

> La constancia de CI es la suposición básica de todas las escalas donde grados relativos de inteligencia se definen en términos de dicho CI. No sólo es básico, sino absolutamente necesario, que los CI sean independientes de la edad en la cual se calculan, debido a que, a menos de que dicha suposición se mantenga, no es posible un esquema permanente de clasificación de la inteligencia.

Aunque los autores contemporáneos de pruebas aceptan, en general, la perspectiva de Wechsler, es importante destacar que la suposición de la invarianza del CI con respecto a la edad es, en realidad, una afirmación de valores, una elección filosófica, y no necesariamente una característica inherente de la naturaleza humana.

Wechsler también tenía la esperanza de utilizar su prueba como un auxiliar para el diagnóstico psiquiátrico. Para lograr este objetivo, dividió su escala en secciones independientes, verbal y de ejecución, lo que permitía que el examinador comparara la facilidad del individuo en el uso de palabras y símbolos (subpruebas verbales) contra su capacidad para manipular objetos y percibir patrones visuales (subpruebas de ejecución). Se consideraba que las grandes diferencias entre capacidad verbal (*V*) y capacidad de ejecución (*E*) tenían importancia diagnóstica. De manera específica, Wechsler consideraba que la enfermedad orgánica del cerebro, la psicosis y los trastornos emocionales, daban lugar a un notable patrón **V>E**, mientras que la psicopatía en adolescentes y el retraso mental leve producían un fuerte patrón **E >V**. La investigación subsecuente demostró muchas excepciones a estas reglas diagnósticas simples. Sin embargo, la distinción entre habilidades verbales y de ejecución ha resultado válida y útil para otros propósitos, como el análisis de las relaciones cerebro-conducta y el estudio de los efectos del envejecimiento sobre la inteligencia. La división teórica de Wechsler para las subpruebas en secciones verbal y de ejecución se clasifica quizá como su contribución más duradera al examen contemporáneo de la inteligencia.

CARACTERÍSTICAS GENERALES DE LAS PRUEBAS WECHSLER

Si se incluyen las revisiones, David Wechsler y sus seguidores produjeron 10 pruebas de inteligencia en un lapso de aproximadamente 60 años. Una de las principales razones para el éxito de estos instrumentos fue que cada nueva prueba o revisión permaneció fiel al contenido y formato familiares que se introdujeron en la Wechsler-Bellevue. Al apegarse a una sola fórmula exitosa, Wechsler aseguró que los examinadores pudieran cambiar de una prueba Wechsler a otra con un reentrenamiento mínimo; esto no sólo fue bueno en términos psicométricos, sino que también representó una hábil estrategia de

mercadeo, en el sentido de que garantizó varias generaciones de usuarios de las pruebas.

Las diversas versiones y ediciones de las pruebas Wechsler poseen las siguientes características en común:

- De 10 a 14 subpruebas. El enfoque de subpruebas múltiples permite que el examinador analice las fortalezas y debilidades intraindividuales, en lugar de calcular una sola puntuación global. Como el lector verá más adelante, el patrón de puntuaciones de subprueba puede transmitir información útil que no es evidente en el nivel global de desempeño.
- Una Escala Verbal compuesta de 5 o 6 subpruebas y una Escala de Ejecución que también incluye 5 o 6 subpruebas. Con esta división, el examinador puede evaluar las habilidades de comprensión verbal y organización perceptual por separado. El patrón de capacidades en estos dos factores de inteligencia puede tener relación con la integridad funcional de los hemisferios izquierdo y derecho del cerebro, y puede servir como indicación de las fortalezas y debilidades vocacionales, como se analiza más adelante.
- Una métrica común para el CI y la puntuación Índice. La media del CI y de la puntuación Índice es de 100 y la desviación estándar es de 15 para todas las pruebas y todos los grupos por edad. Además, las puntuaciones escalares en cada subprueba tienen una media de 10 y una desviación estándar de aproximadamente 3, lo cual permite que el examinador analice las puntuaciones de subprueba de la persona evaluada en cuanto a fortalezas y debilidades relativas.
- Subpruebas comunes para diferentes edades. Por ejemplo, las pruebas preescolar, infantil y adulta de Wechsler (WPPSI-R, WISC-III y WAIS-III) contienen, todas, un núcleo común con las mismas ocho subpruebas (cuadro 6–1). Un examinador que domine la aplicación de una subprueba esencial en cualesquiera de las

Cuadro 6–1. Composición de subpruebas de las escalas Wechsler

	WPPSI-R	WISC-III	WAIS-III
Escalas Verbales			
Información	X	X	X
Retención de Dígitos		X	X
Vocabulario	X	X	X
Aritmética	X	X	X
Comprensión	X	X	X
Semejanzas	X	X	X
Frases	X		
Sucesión de Letras y Números			X
Escalas de Ejecución			
Figuras Incompletas	X	X	X
Ordenación de Dibujos		X	X
Diseño con cubos	X	X	X
Matrices			X
Composición de objetos	X	X	X
Claves/Símbolos y Dígitos		X	X
Laberintos	X	X	
Diseños Geométricos	X		
Búsqueda de Símbolos		X	X
Casa de los animales	X		

Nota: Las subpruebas "principales", comunes a todas las escalas Wechsler, se encuentran en tipografía negrita.

pruebas Wechsler (como la subprueba de Información en el WAIS-III) puede transferir con facilidad esta habilidad a otros miembros de la familia Wechsler de medidas intelectuales.

SUBPRUEBAS WECHSLER: DESCRIPCIÓN Y ANÁLISIS

Wechsler (1939) definió la inteligencia como "la capacidad total o global del individuo para actuar de manera propositiva, pensar de manera racional y manejar de manera eficaz su ambiente". También creía que sólo se puede conocer la inteligencia a través de aquello que le permite hacer a una persona. Por tanto, en el diseño de sus pruebas, seleccionó componentes que representaran un amplio conjunto de capacidades subyacentes, de modo que se pudiera estimar la capacidad intelectual total. Lo que es más, le pidió a sus pacientes que realizaran cosas, no sólo que respondieran a preguntas. Las subpruebas Wechsler son muy diversas y con frecuencia dependen de lo que Wechsler llamaba "producciones mentales".

Se presenta aquí una descripción de las subpruebas del WISC-III y del WAIS-III. También se analizan las capacidades detectadas por cada subprueba y se ofrecen comentarios basados en la investigación. Se refiere al lector al Tema 5B para una descripción de las tres subpruebas únicas del WPPSI-R (Frases, Diseños Geométricos y Casa de los Animales). Las subpruebas verbales se enumeran en primer término.

INFORMACIÓN

En esta subprueba se examina el conocimiento real acerca de personas, lugares y fenómenos comunes. Las preguntas para niños son como las siguientes:

"¿Cuántos ojos tienes?"
"¿Quién inventó el teléfono?"
"¿Qué produce un eclipse solar?"
"¿Cuál es el planeta más grande?"

Las preguntas para adultos son similares, pero progresan a mayores niveles de dificultad. Las preguntas difíciles en la subprueba de Información para adultos se asemejan a:

"¿Cuál es el elemento más común en el aire?"
"¿Cuál es la población mundial?"
"¿Cómo se convierte el jugo de fruta en vino?"
"¿Quién escribió Madame Bovary?"

Los reactivos de prueba de Información examinan los conocimientos generales disponibles para la mayoría de las personas criadas en instituciones culturales y sistemas educativos de las naciones occidentales industrializadas. De manera indirecta, esta subprueba mide el aprendizaje y las habilidades de memoria en el sentido de que los sujetos deben retener el conocimiento obtenido de las oportunidades educativas, formales e informales, a fin de responder a los reactivos de Información.

Por lo común, Información se considera como una de las mejores medidas de capacidad general entre las subpruebas Wechsler (Kaufman, McLean y Reynolds, 1988). Por ejemplo, el manual del WAIS-III señala que, por lo general, Información tiene la segunda o tercera correlación más alta con el CI Total entre todos los 13 grupos por edad (Tulsky, Zhu y Ledbetter, 1997). De manera consistente, Información muestra fuertes cargas en el primer factor identificado en los análisis factoriales de las correlaciones de subprueba del WAIS-III (véase después). Con frecuencia, el primer factor se denomina Comprensión Verbal. Sin embargo, Información tiende a reflejar la educación formal y la motivación para el logro académico y, por tanto, puede producir estimados falsos de capacidad elevada en el caso de estudiantes perpetuos y lectores ávidos.

RETENCIÓN DE DÍGITOS

Retención de Dígitos consta de dos pruebas independientes, Dígitos en Orden Progresivo y Dígitos en Orden Inverso. En la primera, el exa-

minador lee una serie de números a una velocidad de uno por segundo y después le pide al sujeto que los repita. Si el paciente responde correctamente en dos ensayos consecutivos de la misma extensión, el examinador prosigue a la siguiente serie que tiene un dígito adicional, hasta una extensión máxima de nueve dígitos. Para Dígitos en Orden Inverso se utiliza un procedimiento similar, sólo que la persona examinada debe repetir los dígitos en orden inverso, hasta una extensión máxima de ocho dígitos. Por ejemplo, el examinador lee:

"6–1–3–4–2–8–5"

y la persona trata de repetir los números en orden inverso:

"5–8–2–4–3–1–6"

Retención de Dígitos es una medida del recuerdo auditivo inmediato de números. Se requieren facilidad numérica, buena atención y libertad de la distractibilidad. La ansiedad o la fatiga pueden afectar el desempeño en esta subprueba y muchos clínicos han observado que, con frecuencia, los pacientes hospitalizados por razones médicas o psiquiátricas tienen un desempeño deficiente en Retención de Dígitos.

Ambas pruebas pueden evaluar capacidades fundamentalmente diferentes. Dígitos en Orden Progresivo parece requerir que el individuo evaluado acceda en forma secuencial a un código auditivo. En contraste, para el desempeño en Dígitos en Orden Inverso, la persona debe formar una huella visual de memoria a nivel interno a partir de las secuencias numéricas presentadas de modo oral y después recorrer visualmente los números del final al principio. Ésta es, con toda claridad, la prueba más compleja, y no es de sorprender que tenga una carga más fuerte en inteligencia general que Dígitos en Orden Progresivo (Jensen y Osborne, 1979). Gardner (1981) discute que los examinadores deberían complementar los procedimientos estándar de informe y enumerar subpuntuaciones independientes para Retención de Dígitos. Este autor presenta medias, desviaciones estándar y rangos percentiles independientes para Dígitos en Orden Progresivo y Dígitos en Orden Inverso para niños entre 5 y 15 años de edad.

VOCABULARIO

Se pide al individuo que defina hasta varias docenas de palabras que tienen un nivel cada vez mayor de dificultad, mientras el examinador escribe la respuesta al pie de la letra. Por ejemplo, en un reactivo fácil el examinador podría preguntar "¿qué es una taza?", y la persona evaluada obtendría crédito parcial por responder: "una cosa para beber", y crédito completo por responder "tiene un asa, contiene líquido en su interior y bebes de ella". Para los adultos y niños inteligentes, los reactivos avanzados en la subprueba de Vocabulario de Wechsler pueden ser muy desafiantes, al nivel de tintura, vocinglero y atroz.

El vocabulario se aprende en gran medida en el contexto de la lectura de libros y al escuchar a otras personas. Son pocos los individuos que obtienen su vocabulario de la lectura del diccionario o de memorizar las listas de palabras de la sección de "Aumente sus conocimientos de palabras" de las revistas populares. En términos generales, el vocabulario de una persona es una medida de la sensibilidad a información nueva y de la capacidad para descifrar significados a partir del contexto en el que se encuentran las palabras. Precisamente debido a que la adquisición del significado de una palabra depende de la inferencia contextual, la subprueba de Vocabulario resulta ser la mejor medida de la inteligencia general entre las escalas Wechsler (Gregory, 1999). Esto sorprende a muchos legos, quienes consideran que el vocabulario es sólo un sinónimo de la exposición a la educación y, por tanto, un índice mediocre de la inteligencia general. Sin embargo, simplemente no es posible negar la evidencia empírica: Vocabulario tiene la correlación más alta con el CI Total entre todas las subpruebas, tanto en el WISC-III (combinación de los grupos por edad) como en el WAIS-III (para grupos de 12 a 13 años de edad).

ARITMÉTICA

Excepto por los reactivos más fáciles para niños pequeños o personas con retraso, la subprueba de Aritmética está formada por problemas matemáticos presentados de manera oral. El individuo examinado debe resolver los problemas sin papel o lápiz dentro de un límite de tiempo (por lo general 30 o 60 segundos). Los reactivos sencillos destacan operaciones fundamentales de suma o resta, por ejemplo:

"Si tienes 15 manzanas y regalas 7, ¿cuántas te quedan?"

Los reactivos más difíciles requieren una adecuada conceptuación del problema y la aplicación de dos operaciones aritméticas, por ejemplo:

"John compró un equipo estereofónico que tenía una rebaja de 15 % con respecto al precio original de venta de $600. ¿Cuánto pagó John por el equipo estereofónico?"

Aunque los requisitos matemáticos de los reactivos de Aritmética no son excesivamente demandantes, la necesidad de resolver los problemas a nivel mental dentro de un límite de tiempo hace que esta subprueba sea muy desafiante para la mayoría de las personas examinadas. Además de las habilidades aritméticas rudimentarias, el desempeño exitoso en Aritmética requiere altos niveles de concentración y la capacidad para conservar los cálculos intermedios dentro de la memoria a corto plazo. En los análisis factoriales del WISC-III y del WAIS-III, con frecuencia, Aritmética tiene cargas en un tercer factor interpretado ya sea como Libertad de la Distractibilidad o como Memoria.

COMPRENSIÓN

La subprueba de Comprensión es un conjunto ecléctico de reactivos que requieren explicación más que conocimiento sobre hechos reales. Las preguntas fáciles destacan sentido común, mientras las más difíciles requieren una comprensión de los convencionalismos sociales y culturales. En el WAIS-III, dos de las preguntas más difíciles necesitan que la persona examinada interprete proverbios:

Un reactivo fácil de Comprensión se presenta en la siguiente forma: "¿Por qué las personas utilizan ropa?". Los reactivos difíciles se asemejan a los siguientes:

"¿Qué quiere decir este dicho?: 'Más vale pájaro en mano que ciento volando'."

"¿Por qué los jueces del Tribunal Superior de Justicia conservan su puesto de por vida?"

Comprensión parecería ser, en parte, una medida de la "inteligencia social", en cuanto a que muchos reactivos se dirigen a la comprensión de la persona examinada acerca de los convencionalismos sociales y culturales. Sipps, Berry y Lynch (1987) encontraron que las puntuaciones de Comprensión se relacionaban un poco con medidas de la inteligencia social en el Inventario de personalidad de California. Por supuesto, una alta puntuación significa sólo que la persona conoce acerca de los convencionalismos sociales y culturales: la elección de la acción correcta puede provenir o no de este conocimiento. Sin embargo, estudios recientes de Cambell y McCord (1996) y Lipsitz, Dworkin y Erlenmeyer-Kimling (1993) no encontraron sustentación para la creencia popular de que las puntuaciones de Comprensión son sensibles al funcionamiento social.

SEMEJANZAS

En esta subprueba, al individuo examinado se le hacen preguntas del siguiente tipo: "¿En qué se parecen las camisas y los calcetines?" La subprueba de Semejanzas evalúa la capacidad de la persona para distinguir las semejanzas importantes de las no importantes entre objetos, hechos e ideas. De manera indirecta, estas preguntas evalúan la asimilación del concepto de similitud. La persona evaluada debe poseer también la capacidad para juzgar cuándo una semejanza es importante más que trivial. Por ejemplo, las "camisas" y los "calcetines" se

parecen en que ambas palabras comienzan con la letra "c", pero ésta no es la semejanza esencial entre ambos artículos. La similitud importante es que las camisas y los calcetines son, ambos, casos de un mismo concepto, es decir, "ropa". Como ilustra este ejemplo, Semejanzas puede considerarse una prueba de formación de conceptos verbales.

Ahora se centra la atención hacia una descripción y análisis de las subpruebas de Ejecución de las escalas Wechsler. Con excepción de Matrices en el WAIS-III, todas las subpruebas tienen límite de tiempo, y en la mayor parte de ellas se conceden puntos de bonificación a la persona evaluada, con base en un desempeño rápido.

SUCESIÓN DE LETRAS Y NÚMEROS

Ésta es una nueva subprueba que se encuentra sólo en el WAIS-III. El examinador presenta verbalmente una serie de letras y números que se encuentran en orden aleatorio. El examinando debe volver a ordenar y repetir la lista diciendo los números en orden ascendente y después las letras en orden alfabético. Por ejemplo, si el examinador dice "R-3-B-5-Z-1-C", el examinado debe responder "1-3-5-B-C-R-Z." Esta subprueba mide atención, concentración y libertad de la distractibilidad. Junto con Aritmética y Retención de Dígitos, esta subprueba contribuye a la puntuación del Índice de Memoria Simultánea en el WAIS-III (véase después).

FIGURAS INCOMPLETAS

Para esta subprueba, el examinador pide al individuo que identifique la "parte importante" que falta en una ilustración. Por ejemplo, un reactivo simple podría ser de este tipo: un dibujo de una mesa a la que le falta una pata. Los reactivos van aumentando en grado de dificultad; la subprueba continúa hasta que la persona evaluada falla en cinco reactivos consecutivos. La figura 6–1 representa un reactivo similar a aquéllos encontrados en el WAIS-III.



Figura 6–1. Reactivo de Figuras Incompletas similar a los encontrados en el WAIS-III.

Aunque figuras Incompletas está incluida en la mitad de ejecución de cada prueba Wechsler, las capacidades que se requieren para esta subprueba se superponen sólo a un nivel modesto con las medidas típicas de la inteligencia ejecutiva (p. ej., Diseño con cubos). En primer lugar, el desempeño exitoso en Figuras Incompletas implica en gran medida acceder a la memoria a largo plazo, más que utilizar la habilidad perceptual-manipulativa. Es cierto que la persona examinada debe tener buena atención a los detalles visuales, pero las altas puntuaciones reflejan principalmente la capacidad para comparar cada dibujo con objetos o situaciones similares almacenados en la memoria a largo plazo. En resumen, Figuras Incompletas no requiere en realidad de un componente de ejecución. El individuo necesita verbalizar el elemento faltante o sólo señalar a la sección del dibujo que presenta una anomalía. La subprueba de Figuras Incompletas presupone que la persona evaluada se ha visto expuesta al objeto o situación representados, por tal razón, puede ser inapropiada para personas que no han tenido acceso a las ventajas culturales.

ORDENACIÓN DE DIBUJOS

En esta subprueba, el examinador coloca en desorden varios recuadros de una tira de caricaturas no verbales. La tarea de la persona evaluada consiste en colocar los recuadros juntos en el orden correcto, a fin de describir una historia que tenga sentido. La figura 6–2 presenta una tarea de ordenación de dibujos como la que podría encontrarse en el WAIS-III.

Aunque Ordenación de Dibujos se agrupa junto con las tareas de ejecución, tiene cargas aproximadamente iguales en los componentes verbal y de ejecución revelados en los estudios analítico-factoriales de intercorrelación de subpruebas (p. ej., Silverstein, 1982a). Las capacidades detectadas son complejas y multifacéticas. Antes de ordenar los dibujos, la persona examinada debe tener la capacidad de descifrar la gestalt de la historia correcta, a partir de sus elementos desordenados. Esta subprueba también mide pensamiento secuencial y la capacidad para ver relaciones entre acontecimientos sociales. En el WAIS-III, varias de las historias de Ordenación de Dibujos tienen temas humorísticos. Como resultado, se requiere de refinamiento social y sentido del humor para tener un desempeño exitoso.

DISEÑO CON CUBOS

En la subprueba de Diseño con Cubos, la persona debe reproducir diseños geométricos bidimensionales a través de la rotación y colocación correctas de cubos tridimensionales de colores. Esta subprueba se presenta en el Tema 2B, Proceso de prueba. En todas las escalas Wechsler, los primeros reactivos de Diseño con Cubos pueden resolverse mediante tanteo. Sin embargo, los reactivos más difíciles necesitan del análisis de relaciones espaciales, coordinación visomotora y aplicación rígida de la lógica. Diseño con Cubos demanda mucha mayor capacidad de solución de problemas y de razonamiento que la mayor parte de las subpruebas de ejecución, donde la memoria y la experiencia previa tienen mayor ponderación. En los análisis factoriales de las escalas Wechsler, Diseño con Cubos tiene por lo común la carga más elevada de todas las subpruebas de ejecución en el segundo factor. Éste se identifica de diversas maneras como factor no verbal, visoespacial o inteligencia perceptual-organizacional (Fowler, Zillmer y Macciocchio, 1990; Silverstein, 1982a). En el WISC-III y en el WAIS-III, Diseño con Cubos tiene la correlación más alta con el CI de ejecución para casi todos los grupos de estandarización entre las edades de 6 y 89 años. Por esta razón, en general se reconoce a Diseño con Cubos como el índice quintaesencial de la inteligencia no verbal en las pruebas Wechsler (Gregory, 1999).

Ésta es una prueba en la que tienen gran importancia los límites de tiempo. Considérese la versión del WAIS-R, que consta de 14 diseños

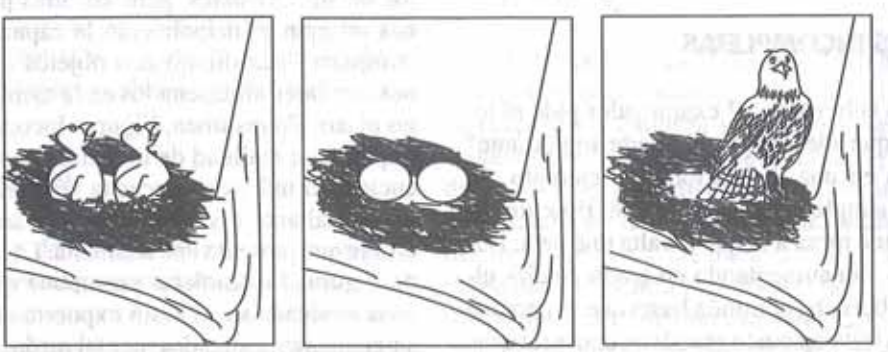


Figura 6–2. Reactivo de Ordenación de Dibujos similar a los encontrados en el WAIS-III.

de dificultad progresiva. Para obtener una alta puntuación, los adultos no sólo deben reproducir correctamente cada uno de los 14 diseños, sino que también deben obtener una bonificación en los últimos cinco diseños al terminarlos con rapidez. Una persona que resuelve todos los diseños dentro del límite de tiempo, pero que no puede obtener los puntos de bonificación, tendrá una puntuación exactamente dentro del promedio en esta subprueba. Las puntuaciones de Diseño con Cubos pueden ser engañosas en el caso de individuos examinados que no valoran la ejecución veloz.

MATRICES

Matrices es una nueva subprueba que se encuentra sólo en el WAIS-III. Se añadió para mejorar la evaluación del razonamiento no verbal en la prueba para adultos. La subprueba consiste en 26 problemas de razonamiento figurativo dispuesto en orden de dificultad creciente (figura 6–3). Para encontrar la respuesta correcta, la persona evaluada tiene que identificar un patrón o relación recurrente entre los estímulos figurativos dibujados a lo largo de una línea recta (reactivos sencillos) o en una rejilla de 3 × 3 (reactivos difíciles) en los que falta el último elemento. Con base en el razonamiento no verbal acerca de patrones y relaciones, la persona debe inferir el estímulo faltante y seleccionarlo de entre cinco opciones que se proporcionan en la parte inferior de la tarjeta.

Matrices se diseñó para constituir una medida de la inteligencia fluida, que es la capacidad para realizar operaciones mentales como la manipulación de símbolos abstractos. Los reactivos detectan integración de patrones, razonamiento por analogía y razonamiento en serie. En general, la subprueba es una excelente medida del razonamiento inductivo con base en estímulos figurativos. Matrices es la única subprueba de ejecución sin límite de tiempo dentro del WAIS-III.

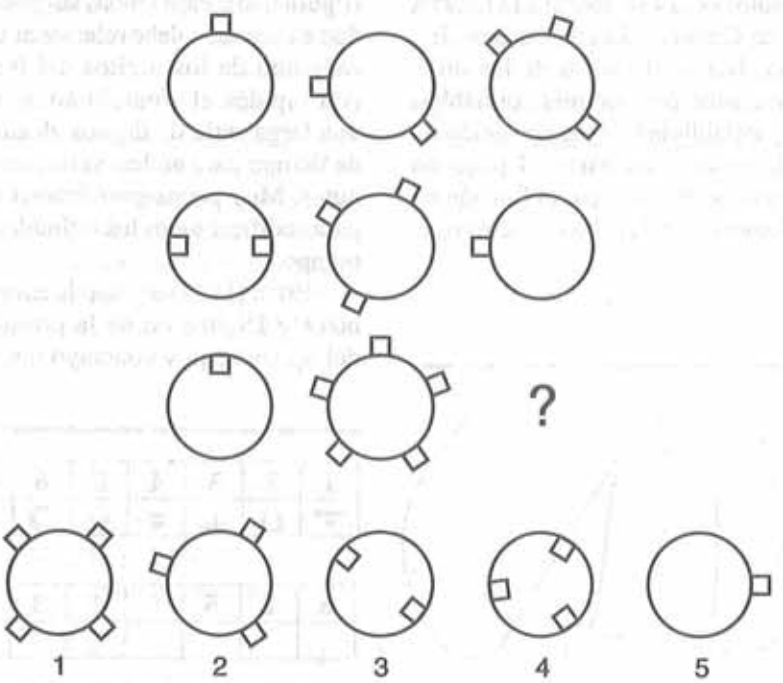


Figura 6–3. Reactivo de Matrices similar a los encontrados en el WAIS-III.

COMPOSICIÓN DE OBJETOS

Para cada reactivo, la persona evaluada debe armar las piezas de un rompecabezas para formar un objeto común (figura 6–4). Por ejemplo, Composición de objetos en el WAIS-III consiste en cinco rompecabezas: un maniquí (seis piezas), un perfil (siete piezas), un elefante (seis piezas), una casa (nueve piezas) y una mariposa (siete piezas). El examinador no identifica los reactivos, de modo que la persona evaluada debe discernir primero el objeto para identificar cada reactivo a partir de sus partes desordenadas. El éxito en esta subprueba requiere altos niveles de organización perceptual; es decir, la persona debe captar el patrón general o gestalt, con base en la percepción de las relaciones entre las partes individuales.

Composición de Objetos es la menos confiable de las subpruebas Wechsler. Por ejemplo, en el WAIS-III esta subprueba tiene una correlación corregida de división en mitades de .70 (Tulsky, Zhu y Ledbetter, 1997). Entre las subpruebas WAIS-III, sólo Ordenación de Dibujos con un valor de .74 se acerca a la falta de confiabilidad de Composición de Objetos. Estas dos subpruebas se destacan de las otras subpruebas Wechsler, que son más confiables. La mediocre confiabilidad de Composición de Objetos puede reflejar, en parte, el pequeño número de reactivos al igual que la función de los factores aleatorios en la solución de rompecabezas.

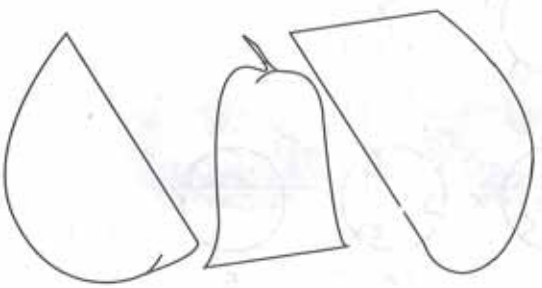


Figura 6–4. Reactivo de Composición de Objetos similar a los encontrados en el WAIS-III.

CLAVES O SÍMBOLOS Y DÍGITOS

Aunque las tareas son casi idénticas, esta subprueba se denomina Claves en el WISC-III y Símbolos y Dígitos en el WAIS-III. La versión del WISC-III consta de dos partes separadas y distintas, una para niños menores a ocho años (Claves A) y otra para niños de ocho años en adelante (Claves B). En Claves A, el niño debe dibujar el símbolo correcto dentro de una serie de formas en secuencia aleatoria. La tarea utiliza cinco formas (estrella, círculo, triángulo, cruz y cuadrado) y a cada forma se le asigna un símbolo único (línea vertical, dos líneas horizontales, una sola línea horizontal, círculo y dos líneas verticales, respectivamente). Después de una breve sesión de práctica, se le pide al niño que dibuje el símbolo correcto dentro de 43 de las formas en secuencia aleatoria. Sin embargo, dado que existe un límite de tiempo de dos minutos, las puntuaciones altas requieren un desempeño rápido.

Claves B en el WISC-III y Símbolos y Dígitos del WAIS-III son idénticas en formato (figura 6–5). Para ambas subpruebas, el individuo examinado debe relacionar un símbolo con cada uno de los dígitos del 0 al 9 y dibujar con rapidez el símbolo apropiado debajo de una larga serie de dígitos aleatorios. El límite de tiempo para ambas versiones es de dos minutos. Muy pocos examinados se las arreglan para codificar todos los estímulos dentro de este tiempo.

Estes (1974) analizó la subprueba de Símbolos y Dígitos desde la postura de la teoría del aprendizaje y concluyó que el desempeño

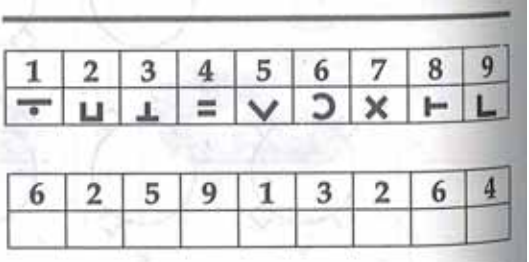


Figura 6–5. Reactivos de Símbolos y Dígitos similares a los encontrados en el WAIS-III.

eficaz requiere la capacidad para producir con rapidez claves verbales para representar cada uno de los símbolos dentro de la memoria. Por ejemplo, en la figura 6–5 la persona examinada podría codificar el símbolo bajo el número dos como "T invertida". La codificación verbal media el desempeño rápido al simplificar la tarea. El desempeño eficiente también demanda aprendizaje inmediato de los pares de símbolos-dígitos, de modo que la persona no tenga que buscar cada dígito en el cuadro de referencia para determinar la respuesta correcta. En este sentido, Símbolos y Dígitos es singular: es la única subprueba Wechsler que necesita de aprendizaje en el momento de una tarea no familiar.

Las puntuaciones de Símbolos y Dígitos muestran una disminución pronunciada a medida que se avanza en edad. En estudios transversales, las puntuaciones naturales en Símbolos y Dígitos declinan hasta 50% de los 20 a los 70 años de edad (Wechsler, 1981). La disminución es aproximadamente lineal y no se explica con facilidad a través de referencias superficiales a las diferencias motivacionales o a la lentificación motriz. Por supuesto, los resultados transversales no son sinónimos por necesidad de las tendencias longitudinales. Sin embargo, la disminución con la edad en Símbolos y Dígitos es tan pronunciada que debe indicar, en parte, un verdadero cambio relacionado con la edad en la velocidad de las habilidades básicas de procesamiento de información. Ésta es una de las subpruebas más sensibles a los efectos del deterioro orgánico (Lezak, 1995).

LABERINTOS

Esta subprueba sólo aparece en el WPPSI-R Y WISC-III, y consiste en laberintos de lápiz y papel que el niño debe resolver dentro de un límite de tiempo. Al individuo evaluado se le pide que no levante el lápiz y se le aconseja que "intente no entrar en callejones sin salida". Se da crédito completo para cada laberinto si el niño lo resuelve dentro del límite de tiempo (30 a 150 segundos, dependiendo de la dificultad) sin entrar en algún callejón sin salida. Se resta un punto de puntuación natural por cada vez que haya entrado en un callejón sin salida.

La subprueba de Laberintos detecta habilidades perceptomotoras, velocidad motora, planificación visual y la capacidad para inhibir las respuestas impulsivas. Esta subprueba es una medida deficiente de la inteligencia general, pero mide razonablemente bien la organización perceptual. En el WISC-III, Laberintos es una subprueba complementaria que no se utiliza para el cálculo del CI.

BÚSQUEDA DE SÍMBOLOS

La Búsqueda de Símbolos es una medida complementaria de ejecución que se encuentra en el WISC-III y el WAIS-III. Es una subprueba con fuerte influencia de los límites de tiempo, en la que el niño observa un grupo blanco de símbolos, después examina con rapidez un grupo de búsqueda con símbolos y finalmente marca la casilla de "SÍ" o "NO" para indicar si uno o más de los símbolos dentro del grupo blanco ocurrieron dentro del grupo de búsqueda. Un reactivo de Búsqueda de Símbolos se representa en la figura 6–6. Esta subprueba parecería ser una medida de la velocidad de procesamiento.



Nota: La tarea de la persona examinada consiste en determinar si cualesquiera de las formas que se encuentran a la izquierda ocurre entre las cinco formas de la derecha.

Figura 6–6. Reactivo de Búsqueda de Símbolos similar a aquellos encontrados en el WISC-III.

ESCALA WECHSLER DE INTELIGENCIA PARA ADULTOS-III

El WAIS-III constituye una revisión significativa del WAIS-R, aunque se conservó la mayor parte de los reactivos anteriores. Los cambios más importantes incluyen la adición de tres subpruebas y de un modelo alternativo para calificar la prueba (cuatro puntuaciones Índice para complementar el abordaje tradicional de CI Verbal, de Ejecución y Total). Otras mejorías importantes con respecto a su antecesor incluyen la actualización y expansión de las muestras normativas, extensión de la cobertura hasta los 89 años de edad, adición de reactivos fáciles para mejorar la evaluación del retraso mental y establecimiento de normas compartidas con la Escala Wechsler de Memoria-III (Gregory, 1999). Debido a los cambios en los protocolos del WAIS-III (p. ej., presentación destacada de reglas de descontinuación), esta prueba es un poco más fácil de aplicar que el WAIS-R. Sattler y Ryan (1999) proporcionan una reseña sobresaliente acerca del WAIS-III en la práctica clínica.

El WAIS-III se compone de 14 subpruebas, pero una (Composición de Objetos) es ahora opcional y se le utiliza sólo como sustituto para pruebas inutilizadas bajo circunstancias poco comunes (Wechsler, 1997). Del conjunto principal de 13 subpruebas, 11 son necesarias para el cálculo de los CI tradicionales (Verbal, Ejecución y Total). Las normas para las puntuaciones de CI se establecieron con el promedio convencional de 100 y desviación estándar de 15 en la población general. La separación de las subpruebas para obtener las puntuaciones de CI es como sigue:

CI Verbal
- Vocabulario
- Semejanzas
- Aritmética
- Retención de Dígitos
- Información
- Comprensión

CI de Ejecución
- Figuras Incompletas
- Claves, Símbolos y Dígitos
- Diseño con Cubos
- Matrices
- Ordenación de Dibujos

Las 11 subpruebas se utilizan para el cálculo del CI Total. La división Verbal-Ejecución de las subpruebas del WAIS-III es casi idéntica a la encontrada en el WAIS-R. La única diferencia es la adición de Matrices en lugar de Composición de Objetos.

Además de las puntuaciones tradicionales de CI, en el WAIS-III se pueden obtener puntuaciones en cuatro Índices, cada uno de los cuales se basa en 2 o 3 de las 13 subpruebas. Estos índices se derivaron del análisis factorial de las subpruebas, el cual reveló cuatro dominios: Comprensión Verbal, Organización Perceptual, Memoria Simultánea y Velocidad de Procesamiento. Las puntuaciones Índice también se basan en la media familiar de 100 y desviación estándar de 15. La división de las subpruebas para las cuatro puntuaciones Índice es la siguiente:

Índice de Comprensión Verbal
- Vocabulario
- Semejanzas
- Información

Índice de Organización Perceptual
- Figuras Incompletas
- Diseño con Cubos
- Matrices

Índice de Memoria Simultánea
- Aritmética
- Retención de Dígitos
- Sucesión de Letras y Números

Índice de Velocidad de Procesamiento
- Claves, Símbolos y Dígitos
- Búsqueda de Símbolos

El lector notará que el Índice de Comprensión Verbal (ICV) es similar al CI Verbal, pero no incluye subpruebas sensibles a la atención (es

decir, Retención de Dígitos y Aritmética). Por tal razón, el ICV es una medida más directa de la comprensión verbal que el CI Verbal. El Índice de Organización Perceptual (IOP) es similar al CI de Ejecución, pero depende menos de la velocidad (porque Matrices no tiene límite de tiempo). Por tal razón, el IOP es una medida más refinada del razonamiento fluido y de la solución de problemas visoespaciales que el CI de Ejecución. En este sentido, el ICV y el IOP son medidas más "puras" que los CI Verbal y de Ejecución, respectivamente.

El Índice de Memoria Simultánea (IMS) se compone de subpruebas sensibles a atención y memoria inmediata (Aritmética, Retención de Dígitos y Sucesión de Letras y Números). Una puntuación relativamente baja en este índice puede significar que la persona evaluada tiene un problema de atención o memoria, en especial con los materiales que se presentan de manera verbal. El Índice de Velocidad de Procesamiento (IVP) se compone de subpruebas que requieren del procesamiento **sumamente** veloz de información visual (Claves, Símbolos y Dígitos, Búsqueda de Símbolos). El IVP es sensible a una amplia variedad de padecimientos neurológicos y neuropsicológicos (Tulsky, Zhu y Ledbetter, 1997).

ESTANDARIZACIÓN DEL WAIS-III

La estandarización del WAIS-III se emprendió con gran cuidado y se basó en los datos obtenidos por el *Bureau of the Census* de EUA en 1995. La muestra total de 2 450 adultos (de 16 y 89 años de edad) se estratificó de manera esmerada en estas variables: sexo, raza, etnicidad, nivel de estudios y región geográfica. Se utilizaron las cifras del censo de 1995 como los valores blanco para las variables de estratificación.

Por ejemplo, entre las personas dentro del rango de 55 a 64 años de edad, el Departamento del Censo encontró que 3.47% son afroestadounidenses con educación media superior. De aquí que el 3.5% de los individuos de estandarización dentro de este rango de edades fueron afroestadounidenses con educación media superior.

La muestra de estandarización se dividió en 13 bandas por edad: 16-17, 18-19, 20-24, 25-29, 30-34, 35-44, 45-54, 55-64, 65-69, 70-74, 75-79, 80-84, 85-89 años. Excepto para los dos grupos de mayor edad, cada muestra incluyó a 200 participantes cuidadosamente estratificados en las variables demográficas que se señalan antes; el grupo de 80 a 84 años de edad incluyó a 150 participantes y el grupo de 85 a 89 años incluyó a 100 participantes. La muestra resultante tiene una correspondencia muy estrecha con las proporciones del censo de EUA. Sin embargo, se excluyó a personas de las que existía la sospecha de un deterioro cognitivo, incluso leve, de modo que es probable que la muestra sea más sana que sus equivalentes del censo. En términos específicos, se utilizaron varios criterios de exclusión dentro de la muestra de estandarización, incluyendo daltonismo, deterioro auditivo o visual sin corrección, evidencia de problemas con drogas/alcohol, menoscabo de extremidades superiores, utilización de medicamentos ansiolíticos o antidepresivos y una variedad de padecimientos potencialmente discapacitantes a nivel neurológico (lesiones craneales, embolias, epilepsia, enfermedad de Alzheimer, esquizofrenia).

Aunque el WAIS-III es muy similar al WAIS-R y tiene una superposición sustancial en cuanto a los reactivos, ambas pruebas no producen CI similares. En estudios contrabalanceados que comparan las puntuaciones de 192 adultos en las dos pruebas, las puntuaciones del WAIS-III son menores en un punto en el caso del CI Verbal, en 5 puntos para el CI de ejecución y en 3 puntos para el CI Total (Tulsky, Zhu y Ledbetter, 1997). En pocas palabras, el WAIS-III es una prueba más difícil que el WAIS-R. Aquí se encuentra un enigma perturbador: ¿por qué la muestra normativa del WAIS-III parece ser más inteligente que la muestra normativa del WAIS-R? Esta cuestión se verá con mayor detalle en el Tema 7B, Predisposición de prueba y otras controversias.

CONFIABILIDAD

La confiabilidad del WAIS-III es excepcionalmente buena. El promedio de las confiabilidades

mixtas de división en mitades entre todos los grupos por edad son: CI Verbal, 0.97; CI de Ejecución, 0.94; y CI Total, 0.98. Los coeficientes de estabilidad en *test-retest* para 394 examinandos confirman en mucho la misma imagen: CI Verbal, 0.96; CI de Ejecución, 0.91; y CI Total, 0.96. Las confiabilidades y coeficientes de estabilidad para las cuatro puntuaciones Índice tienden a ser un tanto inferiores, pero aún se encuentran en o cerca de .90 en todos los casos.

Para el CI Total, el error estándar de medida se encuentra en el rango de 2 a 2.5 puntos, según el grupo de edad. Considérese lo que esto significa: 95% de las veces, el CI Total **verdadero** de una persona evaluada se encontrará dentro de ± 5 puntos (dos errores estándar de medida) del valor obtenido. En términos de uso común, los psicómetras dirían que el CI del WAIS-III tiene una banda de error alrededor de 10 puntos; es decir, las puntuaciones de CI son precisas dentro de un rango de aproximadamente ± 5 puntos.

En contraste con las fuertes confiabilidades encontradas para las puntuaciones de CI e Índice, las confiabilidades de las 14 subpruebas individuales son, en general, mucho más débiles. Las únicas subpruebas con coeficientes de estabilidad superiores a .90 son Información (.94) y Vocabulario (.91). Para las subpruebas restantes, los valores de confiabilidad van desde la parte baja del rango de .70 hasta mediados del rango de .80. La implicación más importante de estas conclusiones sobre confiabilidad débil es que los examinadores deben abordar el análisis de perfil de subpruebas con extremo cuidado. Las puntuaciones de subprueba que parecen tener una elevación (o disminución) discrepante en el caso de un individuo evaluado podrían ser consecuencia de la confiabilidad generalmente débil de ciertas subpruebas, más que una indicación de fortalezas o debilidades cognitivas verdaderas. Algunos estudiosos del tema concluyen que el análisis del perfil (la identificación de fortalezas y debilidades cognitivas específicas con base en el análisis de los picos y valles en las puntuaciones de subprueba) no se justifica con fundamento en la evidencia (Gregory, 1994b).

VALIDEZ

Con base en un número de diferentes líneas de evidencia revisadas aquí, la validez del WAIS-III parece ser bastante satisfactoria. En numerosos estudios que correlacionan el WAIS-III con pruebas establecidas de inteligencia y también con medidas del rendimiento académico, se ha demostrado su buena validez relacionada con el criterio. Por ejemplo, el CI Total del WAIS-III tiene una fuerte correlación con las puntuaciones globales en otras medidas: .93 con el WAIS-R, .88 con el WISC-III (en el caso de los jóvenes de 16 años en los grupos donde se superponen las edades), .64 con Matrices Progresivas Estándar y .88 con la Prueba de Inteligencia Stanford-Binet: Cuarta Edición. El CI del WAIS-III también tiene una fuerte correlación con las ocho subpruebas de la Prueba Wechsler de Rendimiento Individual, al revelar una correlación mediana de .70 (Tulsky, Zhu y Ledbetter, 1997). No existe duda de que el WAIS-R capta los mismos aspectos de la inteligencia global que miden otros instrumentos de amplia utilización.

La validez del WAIS-III también se apuntala en su fuerte superposición con el WAIS-R y con el WAIS original, para los cuales se ha establecido una impresionante cantidad de datos de validez. Para una revisión completa de estos datos, el lector debe consultar a Matarazzo (1972) y a Kaufman (1990). Aquí se presentará un estudio representativo y provocativo en el que se realizó un análisis de correlaciones del lugar académico alcanzado por el alumno y las puntuaciones de prueba de inteligencia. Conry y Plant (1965) correlacionaron las puntuaciones del WAIS con la clasificación de 98 estudiantes de educación media superior en el momento de su graduación. También correlacionaron las puntuaciones del WAIS con el promedio de calificaciones al final del primer año de universidad de 335 estudiantes de una segunda muestra. Los resultados se presentan en el cuadro 6–2. Obsérvese que el CI Verbal pronostica el éxito académico al mismo nivel que el CI Total, mientras el CI de Ejecución conserva una relación más débil con los niveles de

Cuadro 6–2. Correlaciones entre el lugar académico en la secundaria, las calificaciones universitarias y las puntuaciones del WAIS

Subpruebas del WAIS y CI	Educación media superior ($N = 98$)	Universidad ($N = 335$)
Información	0.54	0.48
Comprensión	0.55	0.33
Aritmética	0.45	0.19
Semejanzas	0.50	0.39
Retención de Dígitos	0.37	0.04
Vocabulario	0.65	0.46
Símbolos y Dígitos	0.34	0.15
Figuras Incompletas	0.33	0.20
Diseño con Cubos	0.29	0.19
Ordenación de Dibujos	0.22	0.07
Composición de Objetos	0.17	0.12
CI Verbal	0.63	0.47
CI de Ejecución	0.43	0.24
CI Total	0.62	0.44

Fuente: Adaptado con autorización de Conry, R. y Plant, W. T. (1965). WAIS and group test predictions of an academic success criterion: High School and colege. *Educational and Psychological Measurement, 25*, 493-500.

rendimiento. Obsérvese también que Vocabulario produce la correlación general más elevada de todo el cuadro (0.65) con la posición académica. Estos datos refuerzan en gran medida la inclusión de medidas de vocabulario en las pruebas de inteligencia.

Varios estudios refuerzan la validez de constructo del WAIS-III al mostrar que las puntuaciones de prueba en el caso de diversos grupos son consistentes con la teoría. Por ejemplo, Sattler (1982, 1988) ha señalado que las tendencias relativas a la edad en las subpruebas del WAIS-R (las cuales son muy semejantes a las del WAIS-III) se conforman de manera estrecha a la teoría Cattell-Horn de la inteligencia fluida y cristalizada. El lector recordará del tema anterior que la inteligencia fluida se utiliza para resolver problemas novedosos, mientras la inteligencia cristalizada requiere la recuperación de respuestas aprendidas o habituales. Según la teoría, la inteligencia fluida declina de manera aguda en la vejez, mientras la inteligencia cristalizada permanece constante o aumenta ligeramente (Horn, 1985).

Un análisis del WAIS-R y del WAIS-III indica que las subpruebas verbales dependen mucho más de la inteligencia cristalizada (recuperación de respuestas aprendidas), mientras las subpruebas de ejecución requieren altos niveles de inteligencia fluida (solución de problemas novedosos). De conformidad con las expectativas teóricas, una inspección de los datos normativos revela que las puntuaciones naturales en las subpruebas verbales muestran una disminución mínima a medida que se avanza en edad, mientras las puntuaciones naturales en las subpruebas de ejecución caen de manera vertiginosa en el caso de personas mayores (Wechsler, 1981, 1997). Por supuesto, estos datos son transversales y, por tanto, no constituyen una prueba definitiva de la declinación longitudinal. Sin embargo, los decrementos con la edad en las subpruebas de ejecución son tan notables que ponen a prueba la creencia de atribuirlas a diferencias de cohorte u otros factores artificiales. Es más probable que una proporción significativa de esta disminución sea un descenso verdadero relacionado con la edad que corrobora la teoría Cattell-Horn de la inteligencia.

Otra expectativa consistente con la teoría, que se confirma con los datos empíricos, es una fuerte relación entre los logros educativos y las puntuaciones de CI (Kaufman, 1990). Estas dos variables deberían tener una fuerte correlación con base en dos suposiciones: que la educación eleva la inteligencia y que las personas más inteligentes buscarán en general un nivel más alto de educación.

En apariencia, los análisis de la relación entre las puntuaciones de CI del WAIS-III y los logros educativos todavía no se han terminado.

Sin embargo, la investigación con la edición anterior es relevante para esta cuestión debido a la fuerte semejanza entre los dos instrumentos. Matarazzo y Herman (1984) analizaron el total de años de escolaridad contra el CI Verbal, el CI de Ejecución y el CI Total de los 1 880 individuos utilizados en la muestra de estandarización del WAIS-R. Al excluir a las personas jóvenes, con edades de 16 a 24 años, muchos de los cuales todavía no terminaban su educación, la correlación entre los años de escolaridad terminada y el CI Total fue de 0.63 para los 500 sujetos entre 25 y 44 años de edad y de 0.62 para los 730 individuos entre 45 y 74 años.

Estas conclusiones revelan una correlación muy fuerte entre el logro educativo y las puntuaciones de CI. Por último, el CI de Wechsler y los logros laborales también tienen una fuerte vinculación (Reynolds, Chastain, Kaufman y McLean, 1987), lo cual también brinda un apoyo adicional a la validez del WAIS-III como medida de la inteligencia general.

ESCALA WECHSLER DE INTELIGENCIA PARA EL NIVEL ESCOLAR-III

El WISC se publicó en 1949 como una extensión descendente de la Wechsler-Bellevue original. Aunque se le utilizó ampliamente en los siguientes dos decenios, los psicómetras percibieron varios defectos en ella: ausencia de individuos de raza no blanca en la muestra de estandarización, ambigüedades en la calificación, reactivos inapropiados para niños (p. ej., referencia a "cigarros puros") y ausencia de mujeres y afroestadounidenses en el contenido pictórico de los reactivos. El WISC-R (Wechsler, 1974) y el WISC-III (Wechsler, 1991) corrigieron estos errores.

El WISC-III consta de 10 subpruebas principales y tres complementarias. Las subpruebas verbales y de ejecución se aplican en orden alternado:

Subpruebas Verbales	Subpruebas de Ejecución
Información	Figuras Incompletas
Semejanzas	Claves
Aritmética	Ordenación de Dibujos
Vocabulario	Diseño con Cubos
Comprensión	Composición de Objetos
Retención de Dígitos	Búsqueda de Símbolos
	Laberintos

Retención de Dígitos, Búsqueda de Símbolos y Laberintos son subpruebas complementarias que, por lo común, no se incluyen en el cálculo del CI. Sin embargo, en general, estas subpruebas se aplican debido a la información diagnóstica que proporcionan. En caso de que se interrumpa la aplicación de una subprueba y ésta se arruine o deba omitirse debido a discapacidades especiales, se le puede sustituir con Retención de Dígitos (en el caso de subpruebas verbales) o Laberintos (en las subpruebas de Ejecución). Búsqueda de Símbolos **sólo** puede utilizarse como sustituto de la subprueba de Claves.

La estandarización del WISC-III es excepcionalmente buena, ya que se basó en 100 niños y 100 niñas en cada nivel de edad de los 6 años y medio a los 16 años y medio (N total = 2 200). Estos casos se seleccionaron con cuidado y se estratificaron con base en el censo de EUA para 1988 con respecto a raza/origen étnico (blancos, afroestadounidenses, nativos estadounidenses, esquimales, aleutianos, asiáticos, nativos de las Islas del Pacífico y otros), región geográfica y educación de los padres. Aunque no fue una variable formal de estratificación, el tamaño de la comunidad para la muestra de estandarización del WISC-III se asemeja mucho a los datos del censo. La muestra de estandarización se obtuvo tanto de escuelas públicas como privadas e incluye a niños en programas de educación especial. Un rasgo deseable de la muestra es que 7% de los niños estaba dentro de las categorías de discapacidades para el aprendizaje, trastornos emocionales y discapacidades del habla/lenguaje, etc., y 5% de la muestra incluyó a niños dentro de programas para individuos talentosos. La confiabilidad del WISC-III

es comparable a la del WAIS-R: las tres puntuaciones de CI (Verbal, de Ejecución y Total) muestran confiabilidades de división en mitades y *test-retest* dentro del rango de .90; mientras las subpruebas individuales poseen coeficientes de división en mitades un poco más bajos, que van de .69 (Composición de Objetos) a .87 (Vocabulario y Diseño con Cubos). Las confiabilidades *test-retest* tienden a ser ligeramente más bajas.

La validez del WISC-III depende, en parte, de su superposición con el WISC-R, para el cual se pueden citar docenas de estudios de sustentación. No se desea abrumar con detalles excesivos, de modo que se refiere al lector interesado a Sattler (1988) para una buena revisión de los estudios anteriores. El Manual del WISC-III cita un impresionante conjunto de estudios de validez, que se resumirán aquí. Los estudios preliminares indican fuertes correlaciones con las puntuaciones del WISC-R ($r = .90$ para el CIV; .81 para el CIE y .89 para el CIT), fuertes correlaciones con las puntuaciones del WAIS-R en el caso de personas de 16 años de edad ($r = .90$ para el CIV; .80 para el CIE y .86 para el CIT) y correlaciones un poco menores, pero confirmatorias, con las puntuaciones del WPPSI-R en una muestra de niños de 6 años de edad. Estas correlaciones son casi tan altas como lo permitirían las confiabilidades de las escalas respectivas. Un descubrimiento interesante, que se analiza en el siguiente capítulo, es que los CI del WISC-III son, en promedio, aproximadamente cinco puntos menores que los CI del WISC-R (Vance, Maddux, Fuller y Awadh, 1996).

El WISC-III también muestra correlaciones confirmatorias de la teoría con numerosas pruebas cognitivas, de capacidad y de rendimiento (Wechsler, 1991). Por ejemplo, en un estudio con 27 niños entre los 7 a 14 años de edad a los que se les aplicaron tanto el WISC-III como las *Differential Ability Scales* (Escalas de Capacidad Diferencial, DAS; Elliott, 1990), las puntuaciones del CIV del WISC-III se correlacionaron en .87 con las puntuaciones de Capacidad Verbal del DAS, pero sólo en .58 con las puntuaciones de Razonamiento no Verbal de esa misma prueba. Por el contrario, las puntuaciones del CIE del WISC-III se correlacionaron en .78 con las puntuaciones de Razonamiento no Verbal, pero sólo en .31 con las de Capacidad Verbal. El Manual cita correlaciones consistentes con la teoría —apropiadamente altas para constructos similares, bajas para constructos diferentes— con la *Otis-Lennon School Ability Test* (Prueba de Capacidad Escolar Otis-Lennon), el Test de Retención Visual de Benton-Revisado, subpruebas de la Batería de Pruebas Neuropsicológicas Halstead-Reitan y el *Wide Range Achievement Test-Revised* (Prueba de Aprovechamiento de Amplio Rango-Revisada. Los estudios con grupos especiales de niños —talentosos, con retraso mental, discapacidad para el aprendizaje, hiperactividad, trastornos de conducta, retraso en habla/lenguaje— también proporcionan un fuerte apoyo para la validez del WISC-III.

Los estudios analítico-factoriales de la muestra de estandarización proporcionaron evidencia adicional de la utilidad del WISC-III en la evaluación diagnóstica de niños. Los resultados de numerosos análisis factoriales, incluyendo los análisis separados para cuatro subgrupos según la edad (de 6 a 7, 8 a 9, 11 a 13 y 14 a 16 años de edad) constituyen una fuerte indicación de una solución de cuatro factores. Los primeros dos factores son familiares en estudios previos de las Escalas Wechsler:

Comprensión Verbal	**Organización Perceptual**
Información	Figuras Incompletas
Semejanzas	Ordenación de Dibujos
Vocabulario	Diseño con Cubos
Comprensión	Composición de Objetos

El tercer factor en el WISC-III es un tanto diferente del de su antecesor, y el cuarto factor es nuevo:

Libertad de la Distractibilidad	**Velocidad de Procesamiento**
Aritmética	Claves
Retención de Dígitos	Búsqueda de Símbolos

Una subprueba, Laberintos, reveló una asignación inconsistente a los factores, con cargas débiles en Libertad de la Distractibilidad para niños de 6 a 7 años y débiles en Organización Perceptual para niños de ocho años y mayores.

La solución de cuatro factores se presenta también en las otras muestras, incluyendo pacientes psiquiátricos infantiles (Tupa, Wright y Fristad, 1997), niños canadienses seleccionados de manera aleatoria (Roid y Worrall, 1997) y niños que reciben educación especial (Konold, Kush y Canivez, 1997). En una nota discrepante, Riccio, Cohen, Hall y Ross (1997) encontraron que los factores 3 y 4 no eran importantes desde el punto de vista clínico, es decir, mostraban relación mínima con medidas independientes de atención en una muestra de niños con supuestos problemas de atención.

La solución de cuatro factores para el WISC-III da la posibilidad del informe opcional de Puntuaciones de Índice —similares a las puntuaciones de CI— para cada uno de los cuatro factores. Estas puntuaciones se basan en la media familiar de 100 y desviación estándar de 15. Las Puntuaciones de Índice (Índice de Comprensión verbal, ICV, Índice de Libertad de la Distractibilidad, ILD, etc.) se derivan de la asignación de las subpruebas que se lista antes y sirven para complementar los CI Verbal y de Ejecución. Este análisis factorial del WISC-III es particularmente útil en la evaluación de discapacidades para el aprendizaje y trastornos relacionados. En particular, los niños con discapacidad para el aprendizaje tienen la probabilidad de obtener puntuaciones bajas en el tercer factor, lo cual sugiere que la distracción puede subyacer a algunas formas de este trastorno (Farnham-Diggory, 1978; Kaufman, 1979; Smith, 1983). Parker y Atkinson (1994) proporcionan una fundamentación y métodos específicos para el cálculo preciso (por el contrario de la mera estimación) de las puntuaciones factoriales del WISC-III.

ESCALA DE INTELIGENCIA STANFORD-BINET: CUARTA EDICIÓN

Con una herencia que data de la escala Binet-Simon de 1905, la Stanford-Binet:Cuarta Edición (SB:CE) tiene el linaje más antiguo, y quizá más prestigioso, de cualquier prueba de

Cuadro 6–3. Hitos en el desarrollo de la Stanford-Binet y sus pruebas antecesoras

Año	Pruebas/Autores	Comentario
1905	Binet y Simon	Prueba simple de 30 reactivos
1908	Binet y Simon	Introducción del concepto de edad mental
1911	Binet y Simon	Extensión para incluir a los adultos
1916	Stanford y Binet Terman y Merrill	Introducción del concepto de CI
1937	Stanford-Binet-2 Terman y Merrill	Primera utilización de formas paralelas (I y M)
1960	Stanford-Binet-3 Terman y Merrill	Utilización de métodos modernos de análisis de reactivos
1972	Stanford-Binet-3 Thorndike	Reestandarización de la SB-3 con 2 100 personas
1986	Stanford-Binet-4 Thorndike, Hagen y Sattler	Reestructuración completa en 15 subpruebas

inteligencia individual. En el cuadro 6–3 se resumen algunos de los hitos importantes en el desarrollo de la Stanford-Binet y sus predecesoras. La SB:CE incorpora las principales revisiones en cuanto a fundamentación teórica y contenido de las pruebas de inteligencia. Aunque se han conservado muchos de los tipos de reactivos de las anteriores escalas Binet-Simon, la SB:CE es, en realidad, una nueva prueba que comparte sólo una modesta superposición con sus antecesoras.

MODELO DE INTELIGENCIA DE LA SB:CE

La más importante de las alteraciones de la SB:CE es un modelo jerárquico de la inteligencia con múltiples subpruebas. En ediciones anteriores, el examinador sólo obtenía una puntuación mixta de CI. Aunque podía analizarse de manera cualitativa el patrón de respuestas correctas e incorrectas, las pruebas anteriores no proporcionaban una base para el análisis cuantitativo de los subcomponentes de toda la escala. Esta deficiencia se corrige en la SB:CE, que está formada por 15 subpruebas diseñadas para evaluar la capacidad en cuatro áreas: verbal, abstracta/visual, cuantitativa y de memoria a largo plazo. Debido a que las diferentes subpruebas son adecuadas para distintas edades, a un individuo sólo se le aplican de 8 a 10 subpruebas. Por tanto, la SB:CE produce hasta 10 puntuaciones de subprueba, cuatro puntuaciones de área y una puntuación mixta (que ya no se denomina CI) basada en toda la prueba.

Los autores de la SB:CE adoptaron un modelo jerárquico de tres niveles de la estructura de las capacidades cognitivas como una guía para la construcción de esta prueba (figura 6–7). El primer nivel es el de capacidad general *g*, que los autores definen como "la unidad cognitiva y los procesos de control que un individuo utiliza en la organización de estrategias adaptativas para la solución de problemas nuevos" (Thorndike, Hagen y Sattler, 1986). Tomando una extensa muestra de un amplio rango de tareas cognitivas, los factores específicos de cualquier subprueba particular disminuyen en importancia y la posición de la persona examinada en el desempeño combinado se vuelve un reflejo indirecto del factor de inteligencia *g*.

El segundo nivel parece reconocer una aceptación de la perspectiva modificada de Cattell-Horn acerca de la inteligencia (p. ej.,

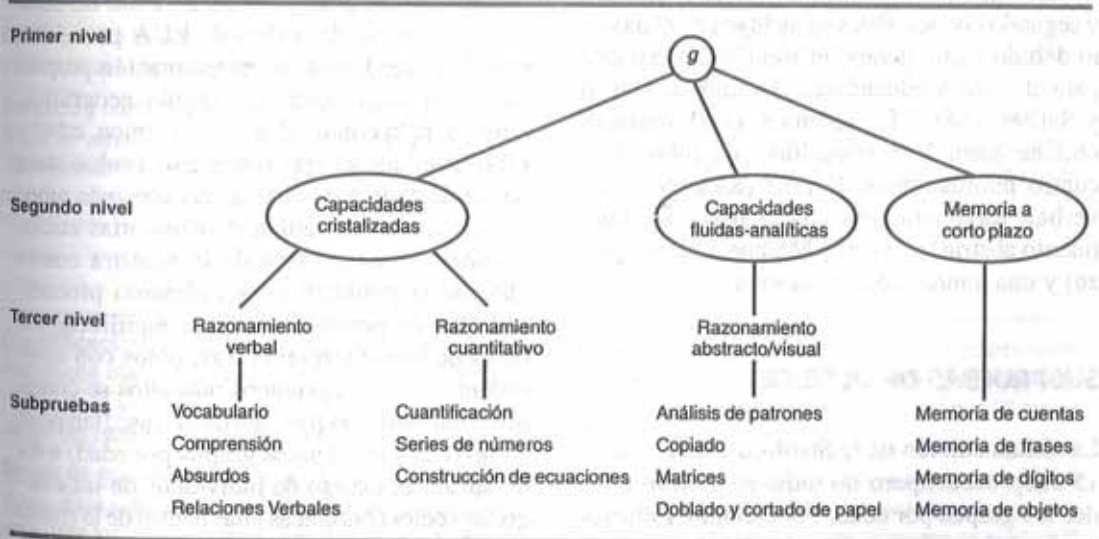


Figura 6–7. Factores de Capacidades Cognitivas y Subpruebas de la Stanford-Binet: Cuarta Edición.

Horn y Cattell, 1966) por parte de los autores de la SB:CE. Sin embargo, además de la distinción entre inteligencia cristalizada/fluida propuesta por Cattell y Horn, los autores han añadido un factor independiente de memoria a corto plazo que se basa en cuatro subpruebas. La SB:CE enfatiza las habilidades de memoria a un grado mucho mayor que cualquier otra prueba de inteligencia general. Cuando el examinador desea una evaluación amplia de la memoria a corto plazo, la SB:CE puede muy bien ser el instrumento de elección. El factor de capacidades cristalizadas representa las habilidades relacionadas con la escuela, como el vocabulario y las habilidades cuantitativas. El factor de las capacidades fluidas-analíticas es, en gran medida, un sinónimo del constructo de Cattell-Horn acerca de la inteligencia fluida y tipifica las habilidades cognitivas necesarias para resolver problemas nuevos que contienen estímulos no verbales y figurativos.

El tercer nivel atañe en realidad sólo al factor de las capacidades cristalizadas, que los autores de la SB:CE dividen en dos áreas: razonamiento verbal y razonamiento cuantitativo. A este nivel, las capacidades fluidas-analíticas reciben la etiqueta alternativa de razonamiento abstracto/visual. Los autores señalan que estos factores son más específicos y más dependientes del contenido que los factores en el primero y segundo niveles. Éstos se incluyen en el modelo debido a que tienen un significado especial para clínicos y educadores (Thorndike, Hagen y Sattler, 1986). En resumen, el examinador obtiene hasta 10 puntuaciones de subprueba, cuatro puntuaciones de área (Razonamiento verbal, Razonamiento cuantitativo, Razonamiento abstracto/visual y Memoria a corto plazo) y una puntuación mixta total.

SUBPRUEBAS DE LA SB:CE

La última edición de la Stanford-Binet incluye 15 subpruebas, pero no todas se aplican a todos los grupos por edad. Por ejemplo, Relaciones Verbales y Construcción de Ecuaciones son demasiado difíciles para niños pequeños y, por lo común, se aplican a personas de ocho años en adelante. En contraste, Absurdos y Copiado son demasiado fáciles para personas mayores y, por lo común, se aplican sólo a individuos menores de 10 años de edad. Seis subpruebas abarcan todos los grupos por edad: Vocabulario, Comprensión, Cuantificación, Análisis de Patrones, Memoria de Cuentas y Memoria de Oraciones.

Nueve de las subpruebas de la SB:CE se basan en los tipos de reactivos que aparecen en ediciones anteriores. Sin embargo, por primera vez los reactivos similares se agrupan en subpruebas. En ediciones previas de la Stanford-Binet, los reactivos se agrupaban por nivel de edad y no por contenido. Seis de las subpruebas de la SB:CE son nuevas y ayudan a corregir el fuerte énfasis verbal de las versiones anteriores. Las nuevas subpruebas proporcionan una cobertura más representativa de las tareas cuantitativas, espaciales y de memoria a corto plazo. Las subpruebas de la SB:CE se describen en el cuadro 6–4.

ESTANDARIZACIÓN DE LA SB:CE Y PROPIEDADES PSICOMÉTRICAS

La muestra de estandarización se aplicó a 5 013 individuos con edades de 2 a 24 años. Se utilizaron los datos del censo de EUA para 1980 con el fin de derivar la representación proporcional en cinco variables: región geográfica, tamaño de la comunidad, grupo étnico, edad y sexo. También se obtuvieron datos sobre clase social. Debido a que había un exceso de niños cuyos padres provenían de categorías socioeconómicas altas (43% de la muestra contra 19% de la población), se utilizaron procedimientos de ponderación para equilibrar los datos de estandarización. Los niños con antecedentes socioeconómicos más altos se contaron como sólo una fracción de un caso (tan poco como 0.28 para algunos grupos por edad) a fin de ajustar el exceso de individuos de las categorías socioeconómicas altas dentro de la muestra. Éste es un procedimiento estadístico razonable y reconocido que mejora los datos

Cuadro 6–4. Descripción de las subpruebas de la Stanford-Binet: Cuarta Edición

Razonamiento verbal

Vocabulario: Comienza con 14 reactivos pictóricos de vocabulario en los que la tarea del niño pequeño consiste en identificar el objeto representado en el dibujo; continúa con 32 palabras que el sujeto debe definir. Las palabras difíciles de vocabulario son similares a:

renuente teratogénico pusilánime

Comprensión: Reactivos sencillos incluyen la identificación de partes del cuerpo humano, mientras que los difíciles requieren juicio social, razonamiento y evaluación. Un ejemplo de un reactivo difícil es: "¿Cuáles son algunas de las razones por las que debe requerirse que un trabajador nuevo se una al sindicato de trabajadores que representa a sus compañeros?"

Absurdos: Esta subprueba se extiende sólo hasta los 9 años de edad y requiere que la persona examinada identifique la incongruencia en las ilustraciones. Esta tarea demanda percepción de los detalles, concentración y comprensión social. Un reactivo característico podría representar a una persona que claramente camina hacia adelante en la arena, con huellas al frente de ella, pero sin ninguna huella detrás.

Relaciones verbales: A la persona examinada se le lee (y también puede ver) una lista de cuatro palabras, con instrucciones de expresar la semejanza entre las primeras tres cosas, pero la diferencia con respecto a una cuarta. Esta subprueba requiere formación de conceptos y razonamiento verbal. Ejemplo: "¿En qué se diferencian una marsopa, un delfín y una ballena de un tiburón?"

Razonamiento abstracto/visual

Análisis de patrones: Los reactivos fáciles implican colocar formas recortadas en los huecos de un tablero de formas, mientras que los reactivos difíciles demandan que la persona reproduzca diseños complejos por medio de cubos. Los reactivos difíciles requieren organización perceptual y visualización espacial, mientras que los reactivos de tablero de formas requieren capacidades visomotoras y de manipulación.

Copiado: En el caso de niños menores de 10 años, el sujeto debe reproducir un modelo simple de cubos sólidos (para individuos muy pequeños) o copiar un diseño con papel y lápiz. Esta subprueba requiere que la persona integre los procesos perceptual y motor.

Matrices: La tarea de la persona consiste en seleccionar el objeto, diseño o letra que complete mejor la matriz. Esta subprueba requiere razonamiento perceptual y tiene fuertes cargas en inteligencia fluida.

Doblado y cortado de papel: Es una prueba de opción múltiple en la que la persona examinada debe identificar cómo se vería un trozo de papel que se ha doblado y cortado si se le desdoblara. Esta subprueba demanda visualización, pensamiento espacial y atención a señales visuales. Doblado y cortado de papel contribuye de manera sustancial al factor de Razonamiento no verbal/Visualización.

Razonamiento cuantitativo

Cuantificación: Las tareas en esta subprueba van desde el conteo simple hasta los conceptos y operaciones aritméticos más completos. Esta subprueba es similar a Aritmética en las escalas Wechsler, pero depende menos de la educación formal. Un ejemplo de un reactivo difícil es: "¿Cuántas losetas de 10 cm por 10 cm se necesitan para cubrir una sección de piso de 1 m por 1 m?"

Series de números: La persona examinada debe llenar una secuencia numérica con el siguiente número lógico. Esta subprueba mide razonamiento lógico y requiere persistencia, flexibilidad y un enfoque de ensayo-error para los problemas difíciles. Un ejemplo de un reactivo difícil es: "¿Cuales son los números siguientes?"

5, 10, 9, 18, 17, 34, 33, ___, ___.

Cuadro 6–4. Descripción de las subpruebas de la Stanford-Binet: Cuarta Edición (continuación)

Construcción de ecuaciones: El sujeto debe reordenar correctamente una ecuación aritmética desordenada. Como en el caso de Series de números, esta subprueba demanda lógica, flexibilidad y persistencia. Un ejemplo de un reactivo difícil es: 1/4 4 6 10 − / =

Memoria a corto plazo

Memoria de cuentas: Utilizando cuatro formas de cuentas de tres diferentes colores (un total de 12 tipos de cuentas), la persona examinada debe reproducir un diseño en fotografía colocando cuentas reales en una estaca. Sin embargo, el diseño está a la vista sólo durante cinco segundos, de modo que la tarea debe llevarse a cabo de memoria. No tiene límite de tiempo.

Memoria de frases: La persona debe repetir al pie de la letra frases presentadas de manera oral. Esta subprueba tiene fuertes cargas en Comprensión verbal a edades de 2 a 7 años, pero es una buena medida del factor de Memoria para sujetos mayores. Ejemplo de una frase difícil es: "A medida que los profesores se hacen más viejos, en general les parece cada vez menos importante publicar muchos artículos breves cada año".

Memoria de dígitos: Esta subprueba es esencialmente idéntica a Retención de Dígitos de las pruebas Wechsler. Mide memoria auditiva a corto plazo y libertad de la distractibilidad. Incluye dígitos en orden progresivo y dígitos en orden inverso.

Memoria de objetos: En esta subprueba, a la persona se le muestra una serie de ilustraciones de objetos individuales a una frecuencia de una por segundo. Después de que se han presentado todas las tarjetas para un reactivo, el individuo debe elegir los objetos mostrados entre aquéllos presentados en una tarjeta que contiene de 5 a 13 elementos. Para obtener un crédito completo, el sujeto debe señalar a los objetos correctos en el mismo orden en que se le mostraron. Esta subprueba contribuye de manera sustancial al factor de Memoria en todos los niveles de edad.

Nota: Estos reactivos se parecen a los de la Stanford-Binet, pero en realidad no se encuentran dentro de la prueba.

de estandarización. En general, parece ser que los datos normativos para la SB:CE son bastante representativos de la población blanco.

Con excepción de Memoria de Objetos, la confiabilidad de las subpruebas de la SB:CE es buena, a juzgar por los coeficientes Kuder-Richardson de consistencia interna dentro de los rangos de .80 y .90 obtenidos con casi todos los grupos por edad. La confiabilidad de Memoria de objetos fue sólo regular, con coeficientes de consistencia interna que por lo general se encuentran en la parte baja del rango de .70. La confiabilidad de las puntuaciones de área es sobresaliente, en especial cuando se utiliza un complemento íntegro de subpruebas para obtener una puntuación de área (los coeficientes de consistencia interna se encuentran en la parte superior del rango de .90). La puntuación mixta es excepcionalmente confiable, con coeficientes KR-20 que van de .95, a la edad de dos años, a .99 a edades de 18 a 23 años. La confiabilidad *test-retest* se evaluó en muestras de niños de 5 y 8 años de edad. Con un intervalo promedio de 16 semanas entre sesiones, las confiabilidades en el *retest* se colocaron dentro del rango de .90 para la puntuación mixta, mientras las puntuaciones de área y de subprueba demostraron un patrón más variable. Por ejemplo, muchas de las confiabilidades en el *retest* se colocaron en los rangos de .60 y .70, que es demasiado bajo para la toma de decisiones acerca de los individuos. En pocos casos, las confiabilidades de subprueba fueron extremadamente bajas: .28 para Cuantificación, .46 para Copiado y .51 para Razonamiento Cuantitativo, en la muestra de niños de ocho años de edad, y .56

para Memoria de Cuentas en el caso de la muestra de niños de cinco años. Los examinadores harían bien en interpretar con precaución las puntuaciones en estas subpruebas.

APLICACIÓN Y CALIFICACIÓN DE LA SB:CE

La SB:CE utiliza la primera subprueba, Vocabulario, como una prueba directiva para el resto del examen. El propósito de una prueba directiva consiste en reducir al mínimo el número total de reactivos aplicados al saltarse los reactivos fáciles de subprueba que, casi con absoluta seguridad, la persona examinada aprobaría de todas formas. Con base en la edad cronológica y la puntuación de Vocabulario, el examinador consulta un cuadro para determinar el nivel de entrada para todas las subpruebas restantes. Además, el examinador debe establecer un nivel basal y uno superior para cada subprueba. En el caso de pruebas como la SB:CE en las que los reactivos de subprueba se califican de los más fáciles a los más difíciles, el nivel basal es el punto bajo en el cual casi con toda seguridad la persona respondería de manera correcta a todas las preguntas. El nivel basal se encuentra cuando se aprueban los cuatro reactivos en dos niveles consecutivos. A veces, el examinador debe dirigirse hacia abajo del nivel de entrada para encontrar el nivel basal. El nivel superior es aquél por encima del cual casi con toda seguridad el individuo fallaría en todos los reactivos restantes. La prueba se descontinúa en el nivel superior, que se encuentra cuando la persona reprueba ya sea en 3 de los 4 o en los 4 reactivos completos en dos niveles consecutivos.

Cada reactivo se aprueba o reprueba de acuerdo con normas muy específicas. Se proporcionan guías extensas de calificación para cinco subpruebas: Vocabulario, Comprensión, Absurdos, Copiado y Relaciones Verbales. Estas subpruebas requieren respuestas "libres" (lo contrario de respuestas de opción múltiple) y puede necesitarse cierta cantidad de juicio para la calificación de las respuestas poco convencionales. Los autores de la SB:CE no informan los estimados de confiabilidad intercalificadores; sin embargo, con base en la experiencia con otras pruebas, es poco frecuente que las diferencias entre examinadores en la interpretación de las pautas de calificación constituyan una fuente común de error de medición (Gregory, 1999).

VALIDEZ DE LA SB:CE

La validez relacionada con el criterio de la SB:CE ha sido el tema de numerosas investigaciones que correlacionan las puntuaciones obtenidas en este instrumento con las puntuaciones de otras pruebas de inteligencia y aprovechamiento. El *Manual Técnico* es inusualmente fértil en este sentido, presenta una multitud de estudios que comparan o correlacionan las puntuaciones de la SB:CE con otras medidas. En general, este instrumento se correlaciona bastante bien con otras pruebas individuales de inteligencia. Lo que es más, el patrón de correlaciones entre las puntuaciones de área de la SB:CE y las subpuntuaciones de otras pruebas es, en general, confirmatorio —las subescalas similares muestran correlaciones más altas que aquellas que son diferentes. Por ejemplo, el CI Verbal de las escalas Wechsler se correlaciona mejor con la puntuación de Comprensión Verbal de la SB:CE (una medida similar) que con la puntuación de Razonamiento Abstracto/Visual de la SB:CE (una medida diferente).

Varios estudios han investigado la validez de constructo de la SB:CE con análisis factoriales confirmatorios diseñados para establecer si la prueba corresponde a la teoría que guió su construcción (Kline, 1989; Laurent, Swerdlik y Ryburn, 1992; Kaplan y Alfonso, 1997; Thorndike, Hagen y Sattler, 1986). En general, los estudios prestan poco apoyo a la existencia de los cuatro factores planteados en la construcción de la prueba (razonamiento verbal, razonamiento cuantitativo, razonamiento abstracto/visual y memoria a corto plazo). Por ejemplo, Kline (1989) concluyó que la solución de dos factores de Sattler (1988) funciona mejor para edades de 2 a 6 años, mientras la solución de

tres factores se adapta mejor a los grupos de mayor edad. Con base en una amplia reseña de estudios de validez, Laurent, Swerdlik y Ryburn (1992) también concluyeron que la SB:CE es una prueba bifactorial para edades de 2 a 6 años (verbal, no verbal) y una prueba trifactorial para edades de siete años y mayores (verbal, no verbal, memoria). Sin embargo, Thorndike (1990) defiende el modelo original de cuatro factores. En parte, el debate se reduce a la elección del análisis factorial "correcto" (McCallum, 1990).

La SB:CE es, sin lugar a dudas, mejor que las ediciones anteriores de esta prueba. Uno de los principales cambios —el desglose del desempeño general en puntuaciones de área y subprueba— es una adición bienvenida y útil al examen de la inteligencia; además, incorpora una buena cantidad de reactivos muy fáciles en la mayor parte de las subescalas. En consecuencia, el examinador obtiene una imagen mucho más precisa del funcionamiento de bajo nivel en niños y adultos jóvenes.* Para la evaluación del retraso mental o el daño cerebral significativo, la SB:CE es superior a casi todos sus competidores, en especial aquéllos como el WAIS-R cuyo CI Total "toca fondo" a mitad en las puntuaciones 50. Otra ventaja de la SB:CE es que cuatro subpruebas evalúan memoria a corto plazo. Cuando el problema de referencia implica memoria a corto plazo, es probable que la SB:CE sea el instrumento a elegir.

Sin embargo, la SB:CE tiene varias deficiencias. El problema más grave es la falta de uniformidad de las puntuaciones mixtas. Las puntuaciones más altas y más bajas posibles fluctúan de manera bastante notable a través de los diversos grupos por edad. Por ejemplo, la puntuación mixta más alta posible para las edades de 2 a 12 años es de 164, pero, después de esa edad, el nivel superior cae hasta 149. Otro problema significativo es la falta de una batería comparable para todos los niveles de edad que cubre esta prueba. A los niños no se les aplican las mismas subpruebas que a los individuos de mayor edad. Otro problema con la SB:CE es la confusión acerca de su estructura factorial. ¿Cuántos factores existen, 2, 3 o 4? Parece haber surgido un consenso de que los cuatro factores propuestos por Thorndike *et al.* (1986) no se sustentan en la evidencia y que la solución de dos factores describe el desempeño en prueba de niños en edad preescolar, mientras la solución de tres factores funciona mejor en el caso de niños de siete años en adelante (Laurent *et al.*, 1992). Otra preocupación es la posibilidad de comparación de los resultados de la SB:CE con otras pruebas de inteligencia. Prewett y Matavich (1994) encontraron que las puntuaciones mixtas de prueba de la SB:CE eran, en promedio, siete puntos más altas a las del CI del WISC-III en una muestra de niños canalizados. Peor aún, la puntuación de Razonamiento Verbal de la SB:CE era 12 puntos mayor, en promedio, que el CI Verbal del WISC-III. La enorme diferencia en puntuaciones verbales es perturbadora porque indica que las dos pruebas darían impresiones diagnósticas sumamente diferentes en casos individuales. Sin embargo, es poco claro si las puntuaciones de la SB:CE son demasiado altas o si las puntuaciones del WISC-III son demasiado bajas (o quizás están en juego ambos efectos).

* Como ha señalado Sattler (1988), es posible que la SB:CE no rinda puntuaciones válidas en el caso de individuos de bajo funcionamiento menores a cuatro años de edad. El "piso" de la prueba parece ser artificialmente alto para personas muy jóvenes. Por ejemplo, un niño de dos años de edad que falle en responder a todas las preguntas ¡todavía podría obtener una puntuación de 95!

PRUEBAS DETROIT DE APTITUD PARA EL APRENDIZAJE-4

Las *Detroit Tests of Learning Aptitude*-4 (DTLA-4; Pruebas Detroit de Aptitud para el Aprendizaje; Hammill, 1999) es una revisión reciente de un instrumento que se publicó por primera vez en 1935. La prueba se aplica de manera individual y está diseñada para niños escolares entre 6 y 17 años de edad. La DTLA-4 consiste en 10 subpruebas que forman la base para el cálculo de 16 puntuaciones mixtas (incluyendo inteligencia general, nivel óptimo) y

14 áreas de capacidad. Las subpruebas se encuentran, en gran medida, dentro de la tradición Binet-Wechsler, aunque existen algunas sorpresas, pues incluye una medida de la capacidad narrativa de Construcción de historias (cuadro 6–5).

La puntuación mixta de Capacidad Mental General se forma con la combinación de las puntuaciones estándar de las 10 subpruebas en la batería. La puntuación mixta de Nivel Óptimo se basa en las cuatro puntuaciones estándar más altas obtenidas por el individuo y se considera que representa el desempeño de la persona bajo circunstancias óptimas. Cada una de las 14 puntuaciones restantes de las combinaciones se deriva de una mezcla de varias subpruebas que se considera que miden un atributo común. Por ejemplo, las subpruebas que implican el conocimiento de palabras y su utilización se combinan para obtener la puntuación mixta Verbal, mientras las subpruebas que no implican lectura, escritura o habla comprenden la puntuación mixta No Verbal. Varias de las puntuaciones mixtas están diseñadas para representar los principales constructos dentro de las teorías actuales de la inteligencia. Además de las puntuaciones mixtas del Nivel Mental General y del Nivel Óptimo, las 14 puntuaciones mixtas restantes de la DTLA-3 son las siguientes:

Verbal	No verbal	(Lingüística)
Atención Incrementada	Atención Reducida	(de Atención)
Motriz Incrementada	Motriz Reducida	(Motora)
Fluida	Cristalizada	(Horn y Cattell)
Simultánea	Sucesiva	(Das)
Asociativa	Cognitiva	(Jensen)
Verbal	Ejecución	(Wechsler)

Las 16 puntuaciones mixtas se basan en la media de 100 y en la desviación estándar de 15 ya familiares. Las 10 subpruebas se normaron para una media de 10 y desviación estándar de 3.

Las puntuaciones mixtas se diseñaron para ofrecer evaluaciones contrastantes, de modo que una diferencia entre puntuaciones pueda ser de importancia diagnóstica. Por ejemplo, un individuo que obtiene una puntuación alta en aptitud de Atención-Reducida pero baja en aptitud de Atención-Incrementada (en el dominio de atención), supuestamente sufre de dificultades para la rememoración inmediata, memoria a corto plazo o concentración enfocada.

La DTLA-4 se estandarizó con 1 350 estudiantes cuyos antecedentes eran muy semejantes a los datos del censo en cuanto a sexo, raza, residencia urbana/rural, ingreso familiar, logros educativos de los padres y área geográfica. La confiabilidad de este instrumento es similar a la de otras pruebas individuales de inteligencia, con coeficientes de consistencia interna que por lo general superan a .80 en el caso de las subpruebas, y a .90 en el caso de las puntuaciones mixtas. Asimismo, los coeficientes *test-retest*

Cuadro 6–5. Descripción breve de las subpruebas de la DTLA-4

Subprueba	Tarea
Palabras opuestas	Proporcionar antónimos (palabras opuestas)
Secuencias de diseños	Discriminar y recordar material gráfico sin sentido
Imitación de oraciones	Repetir oraciones presentadas de manera oral
Letras invertidas	Memoria visual a corto plazo y atención
Construcción de historias	Crear una historia lógica a partir de varias ilustraciones
Reproducción de diseños	Copiar diseños de memoria
Información básica	Conocimiento de información y hechos cotidianos
Relaciones simbólicas	Seleccionar entre una serie de diseños la parte faltante de un diseño anterior
Secuencias de palabras	Repetir una serie de palabras sin relación entre sí
Secuencias de historias	Organizar material pictórico en secuencias significativas

para las subpruebas y puntuaciones mixtas se encuentran en los rangos de .80 y .90. La validez relacionada con el criterio se estableció adecuadamente a través de estudios correlativos con otros instrumentos establecidos, como el WISC-III, K-ABC y la Batería Woodcock-Johnson.

Una preocupación con la DTLA-4 es que la división conceptual en puntuaciones mixtas no se sustenta lo suficiente en la evidencia empírica. Por ejemplo, aunque puede ser cierto que la puntuación mixta Simultánea mida los procesos cognitivos simultáneos propuestos por Das, Kirby y Jarman (1979), existe poco apoyo empírico para apuntalar esta afirmación. Otro problema con el instrumento es que ¡existen más puntuaciones mixtas que subpruebas! Es inevitable que tales puntuaciones tengan fuertes intercorrelaciones, dado que cada subprueba ocurre dentro de varias puntuaciones mixtas. En resumen, la DTLA-4 puede utilizarse como una buena medida de la inteligencia general, pero el empleo de las puntuaciones mixtas con propósitos de planificación psicoeducativa requiere estudios empíricos adicionales. Schmidt (1994) proporciona una amplia revisión de la DTLA-3.

TEST BREVE DE INTELIGENCIA DE KAUFMAN (K-BIT)

Las pruebas de inteligencia analizadas antes y en el tema anterior son medidas excelentes de la capacidad intelectual, pero no dejan de tener defectos. Un problema es el tiempo que se requiere para aplicarlas. Las sesiones de prueba con las Escalas Wechsler, la Batería Kaufman de Evaluación para Niños y la Stanford-Binet pueden durar una hora y no es poco común que se utilicen dos horas si el niño es inteligente y propenso a hablar demasiado. Una segunda desventaja de estas pruebas tan conocidas es la cantidad de entrenamiento que se requiere para aplicarlas. La aplicación adecuada de la mayoría de las pruebas individuales de inteligencia se basa en la suposición de que el examinador tiene un título a nivel avanzado en psicología o en un campo relacionado y que ha tenido amplia experiencia bajo supervisión con los instrumentos en cuestión.

Alan Kaufman respondió a la necesidad de una medida breve de inteligencia, con fines de detección y de aplicación fácil, al desarrollar el Test Breve de Inteligencia de Kaufman (K-BIT, Kaufman y Kaufman, 1990; Kaufman y Wang, 1992). El K-BIT consiste en una sección de Vocabulario y una de Matrices. La primera contiene dos partes: Vocabulario expresivo (nombrar dibujos) y Definiciones (proporcionar una palabra con base en una frase breve y en ortografía parcial). La prueba de Matrices requiere resolver analogías de matrices de 2 × 2 y 3 × 3 utilizando estímulos abstractos.

Las normas de la K-BIT se establecieron para personas entre 4 y 90 años de edad y se puede aplicar en 15 a 30 minutos. En esta prueba se obtienen puntuaciones estándar con media de 100 y *DE* de 15 para Vocabulario, Matrices y la combinación de ambas, llamada Puntuación Mixta de CI. A pesar de la comparabilidad de estas dimensiones de calificación con pruebas de inteligencia muy conocidas, los autores de la K-BIT aclaran que su instrumento no tiene el propósito de sustituir a los enfoques tradicionales (p. ej., WPPSI-R, K-ABC, WISC-III o SB:CE). La K-BIT es principalmente una prueba de detección útil para señalar la necesidad de una evaluación más amplia. La brevedad de este instrumento también lo hace la opción natural para investigación sobre inteligencia.

Los datos de confiabilidad para la K-BIT son excepcionalmente fuertes. La confiabilidad de división en mitades y los coeficientes *test-retest* para una variedad de muestras se colocaron en el rango de .90 para Vocabulario, .80 y .90 para Matrices y .90 para la Puntuación Mixta de CI. La muestra normativa de 2 022 personas estuvo dentro de 1 a 3 puntos porcentuales con respecto a las cifras del censo de EUA para 1990 en cuanto a género, región geográfica, raza u origen étnico y logros educativos de los padres (edades de 4 a 19 años) o de los individuos examinados (edades de 20 años en adelante).

El Manual de la K-BIT presenta datos de validez sumamente fuertes, con base en 20 es-

tudios de correlación. Estos resultados son similares a un estudio reciente de la validez concurrente en el que se compararon los resultados de la K-BIT con las puntuaciones del WAIS-R de 200 individuos canalizados a un centro de evaluación neuropsicológica (Naugle, Chelune y Tucker, 1993). La muestra de pacientes incluyó a personas con trastornos convulsivos, lesiones encefálicas, abuso de sustancias, trastorno psiquiátrico, embolia, demencia y otros padecimientos neurológicos. La heterogeneidad de la muestra de pacientes garantizó un amplio rango de capacidad funcional, un rasgo deseable en un estudio de validación. Aunque las puntuaciones de la K-BIT tendieron a ser aproximadamente cinco veces más elevadas que los datos obtenidos del WAIS-R, las correlaciones entre estos dos instrumentos fueron muy altas y constituyeron una confirmación de la teoría. El CI de Vocabulario (K-BIT) y el CI Verbal (WAIS-R) se correlacionaron en .83, el CI de Matrices (K-BIT) y el CI de Ejecución (WAIS-R) se correlacionaron en .77 y los CI totales de ambos instrumentos tuvieron una sorprendente correlación de .88. En un estudio en el que se compararon las puntuaciones de la K-BIT y del WISC-III de 50 estudiantes canalizados, Prewett (1995) también informó fuertes correlaciones (r = .78 para las puntuaciones totales) y descubrió que las puntuaciones de la K-BIT tendían a ser cinco veces mayores que sus equivalentes en el WISC-III. Canivez (1995) encontró puntuaciones comparables entre la K-BIT y el WISC-III para 137 niños de educación primaria y media y también informó correlaciones muy fuertes entre las dos pruebas, en especial para las puntuaciones totales (r = .87). Eisenstein y Engelhart (1997) encontraron que la K-BIT tenía un buen desempeño en cuanto a estimar los CI de adultos canalizados por trastornos neuropsicológicos, pero Donders (1995) recomienda precaución cuando se utilice la prueba con niños que sufren lesión cerebral. La razón es que las puntuaciones de la K-BIT muestran una relación insignificante con la duración del coma, es decir, la prueba no es un buen índice del estado neuropsicológico de los niños. Aun así, la K-BIT es una medida sobresaliente de detección de la inteligencia general para utilizarse en investigación o cuando las restricciones de tiempo impiden el uso de una medida más larga.

RESUMEN

1. Para propósitos de la estimación de la inteligencia general, cualquier instrumento reconocido, que tenga buenas normas, será suficiente. Sin embargo, cuando el propósito es la evaluación individualizada, los examinadores necesitan considerar las fortalezas y debilidades particulares de los instrumentos potenciales.
2. David Wechsler fue un investigador pragmático que se apropió en gran medida de las pruebas Army Alfa y Army Beta para diseñar muchas de las subpruebas de las diversas escalas Wechsler. Para cada una de sus pruebas de inteligencia, Wechsler utilizó de 10 a 13 subpruebas, donde aproximadamente una mitad son verbales y la otra mitad son de ejecución.
3. La primera prueba de este autor fue la Wechsler-Bellevue, publicada en 1939 y actualizada en 1946. Otras pruebas y sus fechas de revisión son: Escala Wechsler de inteligencia para los niveles preescolar y primario-Revisada (1967, 1989), Escala Wechsler de Inteligencia para el nivel escolar-III (1949, 1974, 1991) y Escala Wechsler de inteligencia para adultos-III (1955, 1981, 1997).
4. Todas las escalas Wechsler utilizan el mismo formato: de 10 a 13 subpruebas divididas casi por igual entre el énfasis verbal y el énfasis en la ejecución; una métrica común para el CI, con media de 100 y desviación estándar de 15; un conjunto común de subpruebas, de modo que los examinadores

puedan transferir con facilidad sus habilidades de aplicación de la prueba de una escala Wechsler a otra.

5. La Escala Wechsler de inteligencia para adultos-III (WAIS-III) es la prueba individual más usada para la inteligencia adulta; tiene excelente confiabilidad, una validez bien establecida y agrega una nueva subprueba: Matrices.
6. Los análisis factoriales de la Escala Wechsler de inteligencia para el nivel escolar-III (WISC-III, entre 6 y 16 años y medio de edad) producen con frecuencia una solución de cuatro factores: Comprensión Verbal (principalmente las subpruebas verbales), Organización Perceptual (principalmente subpruebas de ejecución), Libertad de la Distractibilidad (Aritmética y Retención de dígitos) y Velocidad de Procesamiento (Claves y Búsqueda de Símbolos).
7. La prueba Stanford-Binet: Cuarta Edición (SB:CE) adopta un modelo jerárquico de la inteligencia. El factor general se estima a través de la puntuación total. Los subcomponentes del factor general incluyen un factor de capacidades cristalizadas (razonamiento verbal y razonamiento cuantitativo), un factor de capacidades fluidas-analíticas y un factor de memoria a corto plazo.
8. La SB:CE consta de 15 subpruebas, pero no todas se aplican a todos los grupos por edad. La estandarización de la prueba fue excelente. La SB:CE se correlaciona en gran medida con otras pruebas individuales de inteligencia. Sin embargo, los análisis factoriales confirmatorios de la SB:CE no siempre sustentan el modelo de cuatro factores de la prueba.
9. La SB:CE contiene gran cantidad de reactivos fáciles y, por tanto, es especialmente útil en la evaluación del retraso mental. Un problema con la SB:CE es la falta de uniformidad de las puntuaciones mixtas, que fluctúan de manera amplia entre los grupos por edad. Otro problema es la falta de una batería comparable para todos los niveles de edad que cubre esta prueba.
10. *Detroit Tests of Learning Aptitude*-4 (DTLA-4) está formada por 10 subpruebas que componen la base para el cálculo de 16 combinaciones. La DTLA-4 es una buena medida de la inteligencia general, pero el desglose conceptual en 14 áreas de capacidad necesita sustentación empírica.
11. El Test Breve de Inteligencia de Kaufman (K-BIT) es una prueba de detección de la capacidad intelectual que tiene normas adecuadas y que consiste en Vocabulario y Matrices. Aunque el K-BIT produce un CI mayor en aproximadamente cinco puntos al obtenido con el WAIS-R, la correlación tan elevada entre ambos instrumentos ($r = .88$) constituye un fuerte apoyo para la validez de la K-BIT.

TÉRMINOS Y CONCEPTOS CLAVE

Constancia de CI p. 216
Comprensión verbal p. 231
Organización perceptual p. 231
Libertad de la distractibilidad p. 231
Velocidad de procesamiento p. 231

TEMA 6B
Pruebas grupales de inteligencia

Orígenes y características de las pruebas grupales
Batería Multidimensional de Aptitud (MAB)
Escala del *Shipley Institute of Living* (SILS)
Batería multinivel: Prueba de Capacidades Cognitivas (CogAt)
Prueba de Inteligencia Culturalmente Justa (CFIT)
Matrices Progresivas de Raven (RPM)
Perspectiva sobre las pruebas culturalmente justas
Resumen
Términos y conceptos clave

Una prueba grupal de inteligencia permite evaluar de manera rápida y eficiente a docenas o cientos de personas al mismo tiempo. En este tema se presenta al lector una muestra de pruebas grupales destacadas. Para bien o para mal, el número de pruebas grupales que se comercializan en la actualidad es simplemente sorprendente —se encuentran disponibles veintenas de ellas. En los números recientes del *Mental Measurements Yearbook* (Mitchell, 1985; Conoley y Kramer, 1989, 1992) y de la serie de *Test Critiques* (Keyser y Sweetland, 1984-1988) se presentan varias docenas de registros, además de que cada año se publican instrumentos nuevos. La cobertura amplia de este campo en crecimiento es simplemente imposible. En consecuencia, aquí la información se centra en las interrogantes surgidas de las pruebas grupales y después se revisará un muestrario ecléctico de estos diversos instrumentos.

ORÍGENES Y CARACTERÍSTICAS DE LAS PRUEBAS GRUPALES

ORÍGENES DE LAS PRUEBAS GRUPALES

Las primeras pruebas grupales de inteligencia que tuvieron alguna utilidad se desarrollaron a principios del siglo XX en EUA; sin embargo, sus orígenes pueden ubicarse en los esfuerzos de los psicólogos europeos del siglo XIX. La prueba grupal de inteligencia actual tiene una deuda, en especial con las técnicas de llenado de espacios en blanco que desarrolló Ebbinghaus (1896), en el decenio de 1890. Su prueba consistía en varios pasajes de un texto donde se omitían palabras o partes de ellas, como en el siguiente ejemplo breve:

Caperucita Roja

_____ una vez una pequeña y dulce _____, amada por todos ____ que ____ sus ojos en ella. Su abue___ le regaló una caperuza de seda ____, que llevaba ____ el tiempo. Le llamaban _____ Roja. Un ____, su madre le dijo: "Tu _____ está enferma y débil. _____le este pastel y vino. No te salgas del _____ y no _____ con desconocidos."

La tarea del estudiante consistía en llenar el mayor número de espacios en blanco que fuera posible (en varias selecciones), con un límite de tiempo de cinco minutos; por lo común, una sola persona aplicaba la tarea de llenado a un grupo completo. Sólo en cuatro ocasiones, en varios miles de casos, un estudiante llegó a llenar todos los espacios en blanco. Ebbinghaus empleaba el número total de respuestas correctas como una base para comparar a los individuos en cuanto a su capacidad intelectual (DuBois, 1970).

Años después, el éxito práctico de las escalas Binet inspiró a los psicólogos para desarrollar pruebas de inteligencia que pudieran aplicarse de manera simultánea a grandes números de personas. En el capítulo anterior se ha señalado que la necesidad de someter a prueba a miles de reclutas del ejército para la primera Guerra Mundial inspiró a los psicólogos en EUA (bajo la conducción de Robert M. Yerkes) a lograr rápidos avances en la psicometría y el desarrollo de pruebas. Ocurrieron adelantos paralelos en los sistemas escolares, donde los administradores deseaban tener medios eficaces para examinar y asignar a los alumnos. Sin embargo, la técnica de llenado de espacios en blanco y las preguntas abiertas limitaban gravemente la eficacia de la evaluación. Las pruebas grupales evolucionaron rápidamente a su diseño moderno: el formato de opción múltiple.

DIFERENCIAS ENTRE LAS PRUEBAS GRUPALES E INDIVIDUALES

Las pruebas grupales difieren de las individuales en cinco aspectos:

- Formato de opción múltiple contra formato de preguntas abiertas.
- Calificación mecánica objetiva contra calificación por parte del examinador.
- Aplicación grupal contra individualizada.
- Empleo con fines de detección contra empleo dirigido a la planificación de regularización.
- Muestras enormes de estandarización contra muestras simplemente grandes.

Se analizará cada uno de estos conceptos por separado.

La diferencia más obvia es que las pruebas grupales suelen emplear un formato de opción múltiple. Aunque las primeras pruebas de este tipo utilizaban preguntas abiertas, esta característica se dejó de lado rápidamente debido a las cantidades excesivas de tiempo que se requerían para la calificación. Como un resultado del formato de opción múltiple, las pruebas grupales pueden calificarse de manera rápida y objetiva a través de un instrumento de rastreo óptico conectado a una computadora. La calificación por computadora elimina los errores del examinador y los efectos de halo que pueden ocurrir en la calificación de pruebas individuales. Además, los psicómetras obtienen un acceso casi instantáneo al análisis de reactivos y a los bancos de datos de prueba, de modo que la califi-

cación por computadora promueve el rápido desarrollo y revisión de las pruebas grupales.

Éstas también difieren de las pruebas individuales en el modo de aplicación. En una prueba grupal, el examinador se limita, en gran medida, a la lectura de instrucciones y a la imposición de límites de tiempo; existen ínfimas oportunidades de interacción personal entre quien aplica la prueba y quien la toma. Para la mayoría de los individuos examinados esto no tendrá importancia, pero para algunos —los tímidos, aquellos que se confunden, quienes carecen de motivación— la ausencia de *rapport* con el examinador tendrá resultados desastrosos.

Las pruebas tradicionales de inteligencia sobresalen como auxiliares en el diagnóstico y tratamiento de las dificultades de aprendizaje de los individuos, mientras que las pruebas grupales se utilizan con más frecuencia como métodos masivos de detección en apoyo a la toma de decisiones institucionales. Así, las pruebas grupales pueden utilizarse en los sistemas escolares para "señalar" a los niños que tienen necesidad de regularización o enriquecimiento escolar; en entornos industriales, para identificar a los buenos candidatos para un trabajo específico; o en los ambientes militares, para ayudar a seleccionar a los reclutas que tienen discapacidades mentales.

En general, las pruebas grupales se estandarizan en muestras de enorme tamaño —cientos de miles de personas en vez de sólo unos cuantos cientos de casos seleccionados de manera cuidadosa cuando se trata de pruebas individuales. Por supuesto, nunca debe darse por sentada la idoneidad de una muestra de estandarización. Independientemente de que se utilicen muestras de estandarización de gran tamaño para las pruebas grupales o más pequeñas para las pruebas individuales, es importante determinar el grado en que la muestra es representativa de la población en su conjunto.

VENTAJAS Y DESVENTAJAS DE LAS PRUEBAS GRUPALES

Aunque los pioneros en psicología abrazaron con entusiasmo las pruebas grupales, también reconocieron la naturaleza de esta especie de trato con el Diablo: los psicólogos han vendido el alma del individuo examinado a cambio de los beneficios de las pruebas masivas. Whipple (1910) resumió las ventajas de las pruebas grupales, pero también señaló los peligros potenciales:

> La mayor parte de las pruebas mentales puede aplicarse ya sea a individuos o a grupos. Ambos métodos tienen ventajas y desventajas. El método de grupo tiene, por supuesto, el mérito particular de la economía de tiempo; una clase de 50 o 100 niños puede realizar una prueba en menos de una quincuagésima o centésima parte del tiempo que requeriría la aplicación de la misma prueba de modo individual. En ciertos estudios comparativos; por ejemplo, sobre los efectos de unas vacaciones de una semana sobre la eficiencia mental de los escolares, se vuelve imperativo que todos los individuos realicen la prueba al mismo tiempo. Por otro lado, casi con toda seguridad en todo grupo habrá quienes, por una razón u otra, no pueden seguir las instrucciones o ejecutar la prueba al máximo de su capacidad. El método individual permite que el examinador detecte estos casos y, en general, mediante el ejercicio de la supervisión personal, puede obtenerse, como se señaló antes, información valiosa acerca de la actitud de las personas hacia la prueba.

En suma, las pruebas grupales plantean dos riesgos interrelacionados: 1) algunos individuos examinados tendrán puntuaciones muy por debajo de su capacidad verdadera debido a problemas motivacionales o dificultades para seguir instrucciones y 2) las puntuaciones inválidas no se reconocerán como tales, teniendo consecuencias indeseables para estos individuos atípicos. En realidad, no existe una manera simple en que puedan evitarse por completo estos riesgos, que son parte del intercambio en favor de la eficacia cuando se opta por las pruebas grupales. Sin embargo, es posible reducir al mínimo las consecuencias potencialmente negativas si los examinadores analizan con escepticismo las puntuaciones muy bajas y recomiendan evaluación individual para estos casos.

Volvemos ahora nuestra atención hacia un análisis de varias de las pruebas grupales de inteligencia más destacadas. Se recuerda al lector

que, debido al gran número de estos instrumentos, por necesidad, la revisión será seleccionada. Se presenta una revisión equilibrada de instrumentos viejos, establecidos, y nuevos prometedores para el campo, comenzando con una prueba que intenta cerrar la brecha entre las pruebas individuales y grupales de inteligencia.

BATERÍA MULTIDIMENSIONAL DE APTITUD (MAB)

La *Multidimensional Aptitude Battery*, MAB (Batería Multidimensional de Aptitud; Jackson, 1984a) es una prueba grupal reciente para evaluar inteligencia que se diseñó para ser un equivalente de lápiz y papel del WAIS-R (Krieshok y Harrington, 1985). Como el lector recordará, el WAIS-R es un instrumento muy respetado (ahora reemplazado por el WAIS-III) que se utiliza ampliamente como prueba individual de la inteligencia adulta. Kaufman (1983) señala que el WAIS-R "es el criterio de la inteligencia adulta y ningún otro instrumento se le acerca". Sin embargo, un profesional bien entrenado requiere cerca de una hora y media simplemente para aplicar la prueba de Wechsler para adultos a una sola persona. Debido a que el tiempo profesional es muy solicitado, una evaluación completa con la prueba de inteligencia Wechsler —que incluye aplicación, calificación y redacción del informe— puede costar cientos de dólares. Desde hace tiempo, muchos examinadores han sospechado que una prueba grupal apropiada —con las ventajas adicionales de la calificación objetiva y del informe narrativo por computadora— puede proporcionar una alternativa igualmente válida y mucho menos costosa a las pruebas individuales, para la mayoría, pero no para todas las personas.

ANTECEDENTES Y DESCRIPCIÓN

La MAB se diseñó para producir subpruebas y factores paralelos a los del WAIS-R, pero con el empleo de un formato de opción múltiple que puede calificarse por computadora. El objetivo aparente en el diseño de esta prueba consistió en producir un instrumento que pudiera aplicarse a docenas o cientos de personas con la presencia de un solo examinador (y quizá unos cuantos supervisores de examen) con el mínimo entrenamiento. Además, la MAB se diseñó para obtener puntuaciones de CI que tienen propiedades psicométricas similares a las encontradas en el WAIS-R. La MAB es apropiada para individuos de 16 a 74 años de edad y proporciona 10 puntuaciones de subprueba, así como CI Verbal, de Ejecución y Total.

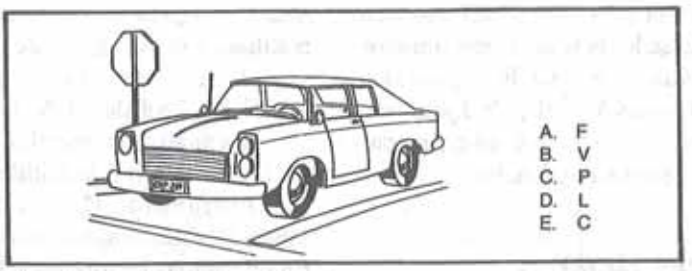
Aunque utiliza reactivos originales de prueba, la MAB es principalmente un clon elaborado del WAIS-R, idéntico en todas las subpruebas. Las 10 subpruebas de la MAB se muestran a continuación:

Verbal	Ejecución
Información	Símbolos y Dígitos
Comprensión	Figuras Incompletas
Aritmética	Espacial
Semejanzas	Ordenación de Dibujos
Vocabulario	Composición de Objetos

El lector observará que en la MAB no se incluye Retención de Dígitos del WAIS-R; la razón es, en gran medida, de naturaleza práctica: simplemente no habría manera de presentar una subprueba parecida en un formato de lápiz y papel; en cualquier caso, la omisión no es grave. Retención de Dígitos tiene la correlación más baja con el CI Total del WAIS-R, y es muy conocido que esta subprueba contribuye poco a la medición de la inteligencia general.

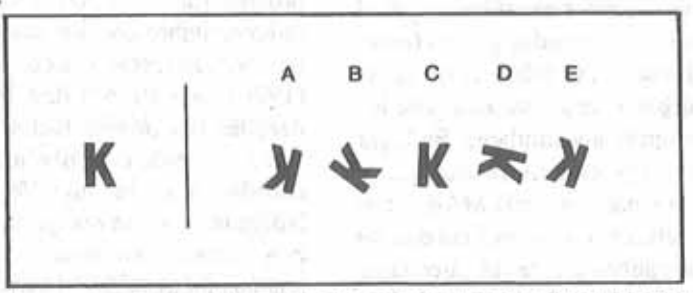
La única diferencia significativa con respecto al WAIS-R es la sustitución de Diseño con Cubos por la subprueba Espacial en la MAB. En ésta, las personas deben ejecutar a nivel mental rotaciones espaciales de figuras y seleccionar como respuesta una de cinco rotaciones posibles presentadas (figura 6–8). Sólo se utilizan rotaciones mentales (aunque se incluyen versiones "invertidas" del estímulo original, como reactivos distractores), los reactivos avanzados son muy complejos y demandantes.

Figuras incompletas: Elija la letra inicial de la palabra que describe la parte faltante en el dibujo



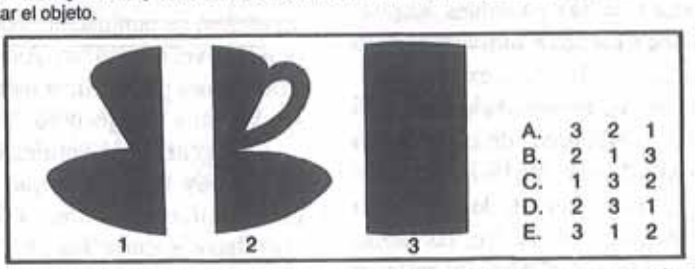
La respuesta es **Faro**, por tanto debe marcarse **A**

Espacial: Elija una figura a la derecha de la línea vertical que sea la misma que la figura a la izquierda. Una de las figuras se puede girar para parecerse a la figura a la izquierda; las otras tendrían que invertirse.



La respuesta correcta es **A**, de modo que ésta debe marcarse. Las otras **BCDE** tendrían que invertirse.

Ensamble de Objetos: Elija el orden, de izquierda a derecha, en que deben colocarse estas partes para formar el objeto.



La respuesta correcta es **C-132**, de modo que se debe marcar **C**. Sólo este orden crearía el objeto **taza**.

Figura 6–8. Reactivos de demostración de tres pruebas de ejecución de la MAB.
Fuente: Reproducido con autorización de Jackson D. N. (1984a). *Manual for the Multidimensional Aptitude Battery.* Port Huron MI: Sigma Assessment Systems, Inc. (800) 265-1285.

Los reactivos en cada una de las 10 subpruebas de la MAB están distribuidos en orden de dificultad creciente; se comienza con preguntas y problemas que la mayoría de los adolescentes y adultos consideran sencillas y continúa con reactivos que son tan difíciles que sólo algunas personas los responden de manera correcta. No se castigan las conjeturas y se alienta a los individuos examinados a responder a todos los reactivos dentro del límite de tiempo. A

diferencia del WAIS-R (donde las subpruebas verbales son medidas de poder sin límite de tiempo), todas las subpruebas de la MAB incorporan elementos tanto de poder como de velocidad: a los individuos sólo se les permiten siete minutos para trabajar en cada subprueba. Incluyendo las instrucciones, las partes Verbal y de Ejecución de la MAB requieren, cada una, aproximadamente 50 minutos para su aplicación.

CARACTERÍSTICAS TÉCNICAS

La primera publicación de la MAB (Jackson, 1984a) no se estandarizó de la manera tradicional, donde las puntuaciones se vinculan con el desempeño de muestras grandes y representativas, estratificadas según variables como sexo, raza, residencia urbana-rural, ocupación de los padres, región geográfica y similares. En lugar de ello, los autores siguieron una estrategia de calibración de las puntuaciones de la MAB con el WAIS-R como prueba de referencia. Para derivar la fórmula lineal de calibración, se aplicaron tanto el WAIS-R como la MAB, en forma contrabalanceada, a una muestra de estudiantes universitarios ($n = 18$), estudiantes del último año de educación media superior ($n = 74$), pacientes psiquiátricos hospitalizados ($n = 58$) e individuos bajo libertad condicional ($n = 10$). Los examinados, 117 varones y 43 mujeres, tenían edades de los 16 a los 35 años. Los coeficientes de correlación entre los CI de la MAB y del WAIS-R fueron de .82, .65 y .91 para los CI Verbal, de Ejecución y Total, respectivamente. De hecho, las tablas normativas presentadas en el *Manual* reflejan una transformación lineal simple de las puntuaciones naturales de la MAB a los CI del WAIS-R para esta muestra inicial de 160 personas.

El *Manual* presenta varios estudios de confiabilidad de consistencia interna y *test-retest*; en general, los resultados son bastante notables. Por ejemplo, en un estudio con más de 500 adolescentes entre 16 y 20 años de edad, la confiabilidad de consistencia interna de los CI Verbal, de Ejecución y Total se colocó en la parte superior del rango de .90. En un estudio *test-retest* con 52 pacientes psiquiátricos jóvenes, las subpruebas individuales mostraron confiabilidades que iban de .83 a .97 (mediana .90) para la escala Verbal, y de .87 a .94 (mediana .91) para la escala de Ejecución. Estos resultados se comparan de manera bastante favorable con los estándares presentados para el WAIS-R (Wechsler, 1981).

Los análisis factoriales de la MAB ofrecen un fuerte apoyo a la validez de constructo de este instrumento (Lee, Wallbrown y Blaha, 1990; Wallbrown, Carmin y Barnett, 1988). En un análisis factorial con las puntuaciones de 3 121 varones y mujeres, estudiantes de bachillerato, el *Manual* presenta un factor general con cargas de moderadas a altas para todas las subpruebas (de .53 a .82). En un análisis factorial independiente con los datos de los individuos de estandarización, Lee, Wallbrown y Blaha (1990) encontraron dos factores ortogonales después del primer factor general. Estos dos factores rotados pueden identificarse con toda claridad como factores Verbal y de Ejecución. Además, otros investigadores han señalado la congruencia extremadamente fuerte entre los análisis factoriales del WAIS-R (sin Retención de Dígitos) y de la MAB. En una muestra de gran tamaño de internos, Ahrens, Evans y Barnett (1990) observaron cambios confirmatorios de validez en las puntuaciones de la MAB en relación con el nivel educativo. Así, existen buenas justificaciones para utilizar las escalas independientes Verbal y de Ejecución en la MAB.

En general, la validez de la MAB depende de su muy fuerte semejanza física y empírica con su prueba madre, el WAIS-R. Los datos correlativos entre las puntuaciones de ambas son importantes en este sentido. Para 145 personas heterogéneas a las que se les aplicaron la MAB y el WAIS-R en forma contrabalanceada, las correlaciones entre ellas fueron de .44 (Espacial/Diseño con Cubos) a .89 (Aritmética y Vocabulario), con una mediana de .78. Las correlaciones entre los CI del WAIS-R y MAB fueron muy saludables; es decir, .92 para el CI Verbal, .79 para el CI de Ejecución y .91 para el CI Total (Jackson, 1984a). Sólo con algunas excepciones, las correlaciones entre las puntuaciones de la MAB y del WAIS-R superan a aquéllas entre el WAIS y el WAIS-R.

COMENTARIO SOBRE LA MAB

Jackson (1984a) tuvo gran cuidado en el desarrollo de la MAB, mejorando de manera continua el instrumento durante un periodo de aproximadamente 10 años antes de publicarlo. Durante este tiempo, los reactivos se seleccionaron, revisaron y eliminaron conforme a criterios psicométricos rigurosos concernientes al nivel de dificultad, poder discriminativo y eficacia de las alternativas distractoras. No es de sorprender que el instrumento resultante sea una hazaña de excelencia psicométrica. Los índices de confiabilidad son fuertes, los análisis factoriales confirman la dicotomía verbal/de ejecución y las puntuaciones de subpruebas y CI generales se correlacionan excepcionalmente bien con las medidas correspondientes del WAIS-R.

A pesar de ello, varios revisores han presentado advertencias y preocupaciones acerca de la MAB que merecen comentario. Krieshok y Harrington (1985) señalan que el *Manual* no proporciona estimados de legibilidad para las instrucciones o para los reactivos en sí, sino que expresa vagamente que la MAB "presupone habilidades de lenguaje necesarias para leer y comprender instrucciones por escrito y para comprender instrucciones verbales". Sin embargo, no recomienda un nivel mínimo de lectura para la aplicación válida; esto puede conducir al examinador a suponer que cualquier persona que llega al nivel mínimo de edad de 16 años puede tomar la MAB, una presuposición que evidentemente es poco firme. De hecho, Krieshok y Harrington (1985) sometieron a la MAB a un análisis de legibilidad por computadora, y concluyeron que algunos reactivos verbales requieren un nivel de lectura del primer año de educación media superior. Debido al nivel relativamente alto de lectura que se necesita en partes de esta prueba, parece probable que un estudiante que en otros sentidos es muy inteligente, pero que tiene discapacidad para la lectura, podría tener una puntuación artificialmente baja en la MAB.

La MAB ha demostrado ser una verdadera promesa en investigación, orientación vocacional y selección de personal; además, puede funcionar como instrumento de detección en entornos clínicos, siempre y cuando el examinador considere las puntuaciones bajas como base para una prueba de seguimiento con el WAIS-R. Los examinadores deben tener en mente que la MAB es una prueba grupal y que, por tanto, lleva consigo el potencial de mal uso en casos individuales. La MAB no debe utilizarse de manera aislada para la toma de decisiones diagnósticas o para asignación a programas como clases para individuos intelectualmente dotados.

ESCALA DEL *SHIPLEY INSTITUTE OF LIVING SCALE* (SILS)

La *Shipley Institute of Living Scale* (SILS) también se conoce como Shipley-Hartford, debido a que se le instituyó hace décadas en Hartford, Connecticut (Shipley, 1940, 1983). La SILS se propuso originalmente como **índice de deterioro intelectual**, en un intento por estimar los efectos de la demencia, daño cerebral y otros padecimientos orgánicos. Sin embargo, la prueba se ha empleado principalmente como una prueba corta de detección de la inteligencia, en especial dentro del sistema de salud mental de la Administración de Veteranos.

ANTECEDENTES Y DESCRIPCIÓN

La SILS se conforma de dos subpruebas, vocabulario y abstracciones. La intención original de la prueba era detectar el deterioro orgánico intelectual al contrastar el desempeño en las secciones de vocabulario y abstracciones. Se consideraba que el deterioro orgánico tenía relativamente poco efecto sobre el vocabulario, y que la capacidad de abstracción mostraría una disminución significativa. Por tanto, una gran discrepancia en favor de vocabulario sobre abstracciones parecería significar la presencia de alteración orgánica. Sin embargo, numerosos estudios y revisiones concluyeron que la SILS tiene un desempeño deficiente como índice del daño orgánico (p. ej., Yates, 1954; Johnson,

1987) por lo que se utiliza pocas veces con este propósito.

La SILS consiste en 40 reactivos de opción múltiple para vocabulario y 20 para pensamiento abstracto; cada reactivo se califica como correcto o incorrecto. Los reactivos abstractos cuentan doble, de modo que la máxima puntuación en cada mitad de la prueba es de 40 puntos. También se presenta una puntuación mixta. La prueba se autoaplica con un límite de tiempo de 10 minutos para cada una de las dos secciones. Algunos usuarios favorecen el uso de la prueba sin límite de tiempo y se han elaborado normas independientes para este enfoque (Heinemann, Harper, Friedman y Whitney, 1985). Pocas personas necesitan más de 10 minutos por sección; la mayoría de los examinadores consideran que la SILS puede ser completamente una medida de poder. También existe una versión de la prueba para microcomputadora; la computadora aplica y califica la prueba y produce un informe narrativo y una representación gráfica de las puntuaciones.

En la sección de vocabulario, la tarea de la persona examinada consiste en seleccionar el sinónimo de una palabra entre cuatro alternativas. Los 40 reactivos se parecen a los siguientes:

• NAVE casa árbol tenedor barco
• PUERIL gordo oportuno necio seco

La puntuación de vocabulario es el número correcto más un punto por cada cuatro reactivos omitidos. Sumar los puntos para los reactivos omitidos proporciona una corrección para la negativa a adivinar. Como resultado de este factor de corrección, la puntuación mínima es aproximadamente de 10 entre 40 puntos.

La intención era que los reactivos de abstracción deberían requerir que la persona examinada obtuviera por inducción algún principio común para una serie dada de componentes y después demostrara su comprensión de este principio terminando la serie. Cada reactivo es una serie de letras o números seguidos de espacios en blanco para indicar el número de caracteres en la respuesta. Los 20 reactivos se parecen a los siguientes:

• A B D G K __
• nao pan arar sara 2 6 8 ___
• 135 341 52 12__

La persona debe terminar cada serie y colocar la respuesta apropiada en los espacios en blanco. (Las respuestas a los reactivos anteriores son P, 962 y 3). Por supuesto, para obtener la respuesta correcta, el individuo debe inferir la regla que determina la progresión de los estímulos en cada reactivo y después utilizar la regla para determinar la continuación. (En el reactivo 1, la distancia entre letras aumenta aritméticamente; en el reactivo 2, los pares son imágenes de espejo una de la otra, excepto por la última y la primera letra que aumentan por uno —de o-p, r-s; en el reactivo 3, cada grupo de números suma uno menos que el grupo anterior - 9, 8, 7, ...).

CARACTERÍSTICAS TÉCNICAS

Zachary (1986) ha publicado normas recientes para la SILS con base en 290 pacientes psiquiátricos diversos que también realizaron el WAIS. La muestra contiene números aproximadamente iguales de varones y mujeres; esta muestra normativa es de jóvenes: la mayoría se encuentran entre 16 y 54 años, con una mediana de edad de 30 años. Con base en esta muestra, el manual contiene tablas de puntuaciones T corregidas según edad (media de 50, *DE* de 10) para vocabulario y abstracciones. Contra el consejo de numerosos investigadores anteriores, el autor del Manual de la SILS también introdujo el Cociente de Abstracción (CA), un nuevo índice de deterioro basado en la diferencia entre las puntuaciones de Vocabulario y Abstracciones. El CA se obtiene al comparar la puntuación pronosticada contra la puntuación obtenida en abstracciones. La puntuación pronosticada se obtiene de una ecuación de regresión que utiliza la puntuación de vocabulario, la edad y el nivel educativo. El CA es una mejoría sobre los índices anteriores de deterioro en cuanto a que, para su cálculo, se toman en cuenta las disminuciones que ocurren de manera natural con la edad. Las personas con esquizofrenia y

otros individuos con disminución en su eficiencia intelectual tienden a obtener bajos CA. Sin embargo, existen causas no patológicas para un CA bajo (p. ej., desagrado por los conceptos abstractos) y, por tanto, la utilidad de este índice está sujeta a dudas.

Mason, Lemmon, Wayne y Schmidt (1991) han intentado revivir el enfoque del CA para la SILS mediante ecuaciones de regresión que se emplean en el cálculo de Cocientes de Abstracción que utilizan la edad, sexo y clase social como variables moderadoras. Sin embargo, no proporcionan ninguna evidencia de la validez del CA como índice de deterioro cerebral.

La confiabilidad de la SILS es marginal. Las medidas típicas de consistencia interna (correlación pares-nones) son de .87 (vocabulario) y .89 (abstracciones) y .92 (puntuación total). Sin embargo, como se señala en los *Standards for Educational and Psychological Testing* (AERA, APA y NCME, 1985), los coeficientes de división en mitades de la variedad pares-nones producen estimados de confiabilidad inflados, cuando se trata de pruebas de velocidad. En el sentido de que las puntuaciones de la SILS se basan en la velocidad en vez del poder, estas confiabilidades serán artificialmente altas. Es más probable que las confiabilidades *test-retest* sean más apropiadas para la SILS. Estas confiabilidades varían de manera considerable en la literatura, pero se acercan al .80 para la puntuación total en muestras más grandes y heterogéneas (Johnson, 1987).

Puesto que la SILS se emplea principalmente como prueba de detección de la inteligencia, la validez de este instrumento se vincula, en gran medida, con su capacidad para pronosticar CI Totales en pruebas individuales como el WAIS o el WAIS-R. Como escribió Johnson (1987), literalmente docenas de estudios de correlación han investigado la precisión de la SILS como medio de predicción del CI Wechsler (p. ej., Zachary, Crumpton y Spiegel, 1985). Las correlaciones entre la SILS y la Wechsler-Bellevue o el WAIS van de .65 a .90, con una mediana de .76 (Johnson, 1987). Con base en estos estudios, Johnson (1987) informa que el intervalo de confianza de 95% para el CI estimado de la SILS es de aproximadamente ± 11 puntos de CI. Por ejemplo, una puntuación total de 60 en la escala Shipley para un varón de 40 años de edad, se convierte en un CI del WAIS-R de 102; en 95% de tales casos, el CI real de la persona en el WAIS-R caerá dentro del rango de 91 a 113 (Zachary, 1986).

COMENTARIO SOBRE LA SILS

La SILS es una medida razonablemente buena de la inteligencia general que ha encontrado amplio uso en la investigación. Además, el instrumento aún es bastante popular como prueba de detección de la inteligencia general y de la posible ineficiencia intelectual (Bowers y Pantle, 1998). Aunque la SILS es útil para la detección intelectual en un rango muy amplio, no se le debe utilizar para realizar discriminaciones más detalladas. Los clínicos responsables utilizarán una prueba de inteligencia individual (p. ej., K-BIT, WAIS-III) cuando se requiera una evaluación individual más precisa.

A pesar de que es una prueba de detección aceptable, la SILS posee varias limitaciones significativas:

1. La SILS es inapropiada para individuos con bajo CI o para los que tienen discapacidades significativas del lenguaje.
2. La prueba tiene un nivel superior bajo, en especial en la sección de abstracciones, y no se distribuye a los individuos examinados que tienen un alto CI.
3. La SILS tiene una banda de error que se aproxima a 11 puntos de CI, lo cual puede ser excesivo para muchas aplicaciones.

BATERÍA MULTINIVEL: PRUEBA DE CAPACIDADES COGNITIVAS (CogAT)

Una función importante de las pruebas psicológicas consiste en evaluar las capacidades de los estudiantes que constituyen un prerrequisito para

el aprendizaje tradicional dentro del salón de clase. En el diseño de pruebas para dicho propósito, los psicómetras deben enfrentarse al problema obvio e irritante de que los jóvenes en edad escolar difieren en gran medida en sus capacidades intelectuales. Por ejemplo, una prueba adecuada para un niño de sexto grado será demasiado fácil para un alumno de secundaria, pero será muy difícil para un estudiante de tercer grado.

La respuesta a este dilema es una batería multinivel, una serie de pruebas superpuestas. En una batería multinivel, cada prueba grupal se diseña para un nivel específico de edad o grado escolar, pero las pruebas adyacentes poseen cierto contenido común. Debido a la superposición en contenido entre los niveles adyacentes de edad o grado, cada prueba posee niveles inferior y superior apropiadamente bajos y altos para la evaluación adecuada de los estudiantes en ambos extremos de capacidad. Además, por lo general, las baterías multinivel emplean muestras normativas sumamente comparables en cada nivel sucesivo. Por todas estas razones, se consideran ideales para estimar la preparación del alumno para el aprendizaje escolar. Casi todos los sistemas escolares en EUA emplean cuando menos una batería multinivel con normas nacionales.

La *Cognitive Abilities Test*, CogAT (Prueba de Capacidades Cognitivas) es una de las mejores baterías de pruebas escolares en uso en la actualidad (Thorndike y Hagen, 1993ab). La revisión más reciente de la prueba es la Edición Multinivel CogAT, Forma 5, publicada en 1993. Este instrumento se analizará con cierto detalle y después se proporcionará un resumen de pruebas contrarias.

ANTECEDENTES Y DESCRIPCIÓN

La CogAT se derivó de las Pruebas de inteligencia Lorge-Thorndike, una de las primeras pruebas grupales de inteligencia que tenía el propósito de utilizarse ampliamente dentro de los sistemas escolares. La CogAT es, en primer término, una medida de la capacidad escolar, pero también incorpora una batería de razonamiento no verbal con reactivos que no tienen relación directa con la educación formal en la escuela. Las dos baterías principales, adecuadas para alumnos desde jardín de niños hasta tercer grado, se analizan de manera breve al final de esta sección. Aquí se revisa la edición multinivel dirigida a estudiantes de 3° al 10° grados.

Las nueve subpruebas de la CogAT Multinivel se agrupan en tres baterías como sigue:

Batería verbal	Batería cuantitativa	Batería no verbal
Razonamiento Verbal	Relaciones Cuantitativas	Clasificación de Figuras
Frases Incompletas	Series de Números	Analogías de Figuras
Analogías Verbales	Construcción de Ecuaciones	Análisis de Figuras

Para cada subprueba de la CogAT, los reactivos se ordenan según el nivel de dificultad en un solo protocolo de prueba. Sin embargo, los niveles de inicio y finalización difieren en cada 1 de los 8 niveles superpuestos (A hasta H). Así, a todos los individuos examinados se les proporcionan los reactivos apropiados según el grado. Todas las subpruebas, excepto una, emplean un formato de opción múltiple. La excepción es Análisis de Figuras, en la que la persona responde con Sí o No a una serie de alternativas.

Las subpruebas tienen estrictos límites de tiempo que van de los 8 a los 12 minutos. Cada una de las tres baterías puede aplicarse en menos de una hora. Sin embargo, el manual recomienda tres sesiones de prueba en días sucesivos para los niños pequeños. En el caso de los niños mayores, el primer día deben aplicarse dos baterías y al día siguiente se utiliza un solo periodo de prueba.

Muchas subpruebas de la CogAT tienen una notable semejanza con partes de la Stanford-Binet: Cuarta Edición. Por ejemplo, ambas pruebas incluyen reactivos de doblado de papel. La explicación reside en la paternidad común: Robert Thorndike y Elizabeth Hagen desarrollaron ambas pruebas y las dos se publicaron en la *Riverside Publishing Company*. De nuevo se ve el carácter híbrido de las pruebas actuales de inteligencia, en cuanto a que las nuevas prue-

bas incorporan los mejores aspectos de sus predecesoras.

Las puntuaciones naturales de cada batería pueden transformarse en puntuación estándar normalizada con base en la edad, con media de 100 y desviación estándar de 16. Además, también están disponibles rangos percentiles y estaninas para los grupos por edad y nivel de grado. Se utilizó una interpolación para determinar las normas de nivel de grado para los periodos de otoño, invierno y primavera.

CARACTERÍSTICAS TÉCNICAS

La CogAT se normó conjuntamente (estandarización concurrente) con tres pruebas de aprovechamiento, las *Iowa Tests of Basic Skills* (Pruebas Iowa de Habilidades Básicas), las *Test of Achievement and Proficiency* (Pruebas de Ejecución y Rendimiento) y las *Iowa Tests of Educational Development* (Pruebas Iowa de Desarrollo Educativo). La estandarización concurrente con medidas de rendimiento es una práctica común y deseable en el establecimiento de normas para pruebas de inteligencia con niveles múltiples. La virtud particular de la normatización conjunta es que se determina con gran precisión la correspondencia esperada entre las puntuaciones de inteligencia y de rendimiento. Como consecuencia, los examinadores pueden identificar con exactitud a los estudiantes de bajo rendimiento que necesitan regularización o de mayor evaluación para descartar una posible discapacidad para el aprendizaje.

La confiabilidad de la CogAT es excepcionalmente buena. En ediciones anteriores, los estimados de confiabilidad Kuder-Richardson 20 para las baterías multinivel promediaron .94 (Verbal), .92 (Cuantitativa) y .93 (No verbal) en todos los niveles por grado. Las confiabilidades *test-retest* a seis meses para las formas alternas fueron de .85 a .93 (Verbal), .78 a .88 (Cuantitativa) y .81 a .89 (No verbal).

El *Manual* proporciona un cúmulo de información acerca de validez de contenido, relacionada con el criterio y de constructo para la CogAT; aquí se resumen sólo los puntos más relevantes. Las correlaciones entre la CogAT y las baterías de rendimiento son sustanciales. Por ejemplo, la batería verbal de la CogAT se correlaciona en los rangos de .70 y .80 con las subpruebas de rendimiento de las pruebas Iowa de habilidades básicas y las de ejecución y rendimiento.

La CogAT pronostica las calificaciones escolares razonablemente bien. Las correlaciones van del rango de .30 al de .60, dependiendo del nivel de grado, sexo y grupo étnico. No parece haber una tendencia clara en cuanto a qué batería es mejor en el pronóstico del promedio de calificaciones. Las correlaciones entre la CogAT y las pruebas individuales de inteligencia también son sustanciales, y por lo común van de .65 a .75. Estos datos hablan bien de la validez de constructo de la CogAT, en la medida en que la Stanford-Binet se reconoce de manera amplia como una excelente medida de la inteligencia individual.

COMENTARIO SOBRE LA CogAT

La edición multinivel de la CogAT es una prueba grupal de la inteligencia con fuerte confiabilidad y normas cuidadosas para estudiantes del tercer a doceavo grados. La estandarización concurrente con dos pruebas de aprovechamiento es un aspecto bienvenido y práctico. En apoyo a la validez de la CogAT, las correlaciones con calificaciones, rendimiento y otras medidas de inteligencia son bastante robustas. Recientemente se ha producido una versión alemana de la CogAT (Perleth, Hofmann, Schauer y Wernberger, 1994).

Ansorge (1985) ha cuestionado si las tres baterías son realmente necesarias. Este autor señala que las correlaciones entre las baterías Verbal, Cuantitativa y No verbal son sustanciales. Los valores medianos en todos los grados son los siguientes:

Verbal y Cuantitativa:	.78
No verbal y Cuantitativa	.78
Verbal y no verbal	.72

Dado que la batería Cuantitativa ofrece pocos aspectos únicos, desde un punto de vista

puramente psicométrico no existe justificación para incluirla. Sin embargo, los autores de la prueba recomiendan utilizar todas las baterías con la esperanza de que las diferencias en desempeño ayudarán a los profesores en la planificación de estrategias de regularización. Sin embargo, los autores no defienden mucho esta afirmación.

Un reciente estudio de Stone (1994) proporciona una notable justificación para la utilización de la CogAT como base para la evaluación de estudiantes. Este investigador encontró que las puntuaciones de la CogAT de 403 alumnos del tercer grado proporcionaban una predicción imparcial del aprovechamiento del alumno que era más precisa que las calificaciones del maestro. Particularmente estas últimas mostraban parcialidad en contra de estudiantes caucásicos y asiaticoestadounidenses al pronosticar menores puntuaciones de aprovechamiento.

PRUEBA DE INTELIGENCIA CULTURALMENTE JUSTA (CFIT)

La *Culture Fair Intelligence Test* (CFIT; Prueba de Inteligencia Culturalmente Justa; Cattell, 1940; IPAT, 1973) es una medida no verbal de la inteligencia fluida que Raymond B. Cattell, el importante psicólogo de la medición, concibió originalmente en el decenio de 1920. El objetivo de la CFIT consiste en medir la inteligencia fluida —capacidad analítica y de razonamiento en situaciones abstractas y novedosas— de una manera que esté tan "libre" del prejuicio cultural como sea posible. Esta prueba se llamó en un principio la Prueba de Inteligencia Culturalmente Libre. El nombre se cambió cuando se hizo evidente la imposibilidad de eliminar por completo las influencias culturales de las pruebas de inteligencia.

ANTECEDENTES Y DESCRIPCIÓN

La CFIT ha atravesado por varias revisiones, surgiendo en su apariencia actual en 1961. La prueba consiste en tres versiones: la Escala 1 se utiliza con adultos con retraso mental y niños entre 4 y 8 años de edad; la Escala 2 se emplea con adultos dentro del rango promedio de inteligencia y con niños entre 8 y 13 años de edad; y la Escala 3 es para adultos con alto nivel de capacidad y para estudiantes de bachillerato y universidad. La Escala 1 implica una considerable interacción entre el examinador y la persona evaluada —cuatro de las subpruebas se deben aplicar de manera individual. Así, en algunos aspectos la Escala 1 es más una prueba individual de inteligencia que una prueba grupal. Aquí se analizan sólo las Escalas 2 y 3, dado que éstas son en realidad pruebas grupales de inteligencia. Estas dos pruebas difieren principalmente en el nivel de dificultad.

Para cada escala se encuentran disponibles dos formatos equivalentes llamados Forma A y Forma B. Los autores de la prueba recomiendan aplicar ambas formas a cada individuo para obtener lo que se denomina prueba completa. Cada forma en sí misma se conoce como forma corta. A pesar de la recomendación de utilizar ambas formas como una prueba combinada, es muy común que los usuarios de la CFIT dependan de una sola forma breve con propósitos de detección.

Cada forma incluye cuatro subpruebas, Series, Clasificación, Matrices y Condiciones. En la figura 6–9 se presentan reactivos de muestra. Por supuesto, cada subprueba está precedida por varios reactivos de práctica. La prueba completa está hábilmente organizada en un protocolo de ocho páginas.

La CFIT es una prueba sumamente rápida. Cada forma de las Escalas 2 y 3 requiere cerca de 30 minutos para su aplicación, pero sólo se dedican 12.5 minutos para la presentación de la prueba en sí. Por tanto, los resultados pueden ser engañosos en el caso de personas que no le dan importancia a la velocidad de ejecución en la resolución de problemas. Por fortuna, la Escala 2 puede utilizarse como una prueba de poder sin límite de tiempo. Sin embargo, las normas para este modo de aplicación son limitadas (IPAT, 1973).

CARACTERÍSTICAS TÉCNICAS

Las muestras de estandarización para las Escalas 2 y 3 fueron respetablemente grandes, pero

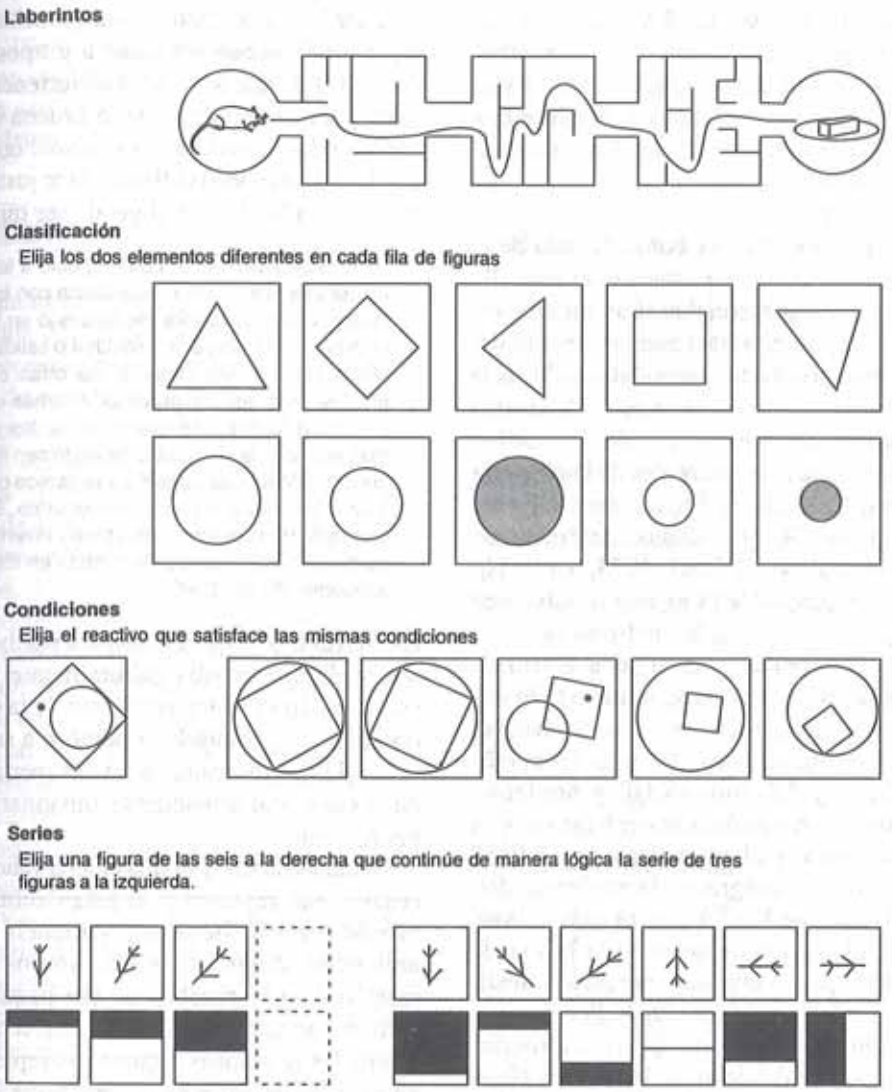


Figura 6–9. Reactivos de muestra de la Prueba de inteligencia culturalmente justa.
Fuente: Copyright © del *Institute for Personality and Ability Testing*, Inc. Reimpreso con autorización.

no se describen con suficiente detalle como para determinar el grado en que reflejan a la población general. Las muestras de estandarización se caracterizan como sigue:

> El grupo de estandarización para la Escala 2 está formado por 4 328 varones y mujeres tomados como muestra en diversas regiones de EUA y Gran Bretaña. Las normas de la Escala 3 se basan en 3 140 casos, que incluyen a alumnos estadounidenses de bachillerato, divididos por igual entre los últimos cuatro años de educación media superior, y jóvenes adultos en una muestra laboral estratificada (IPAT, 1973).

Las puntuaciones naturales se convierten a una puntuación estándar normalizada de CI con media de 100 y desviación estándar de 16.

Las confiabilidades *test-retest*, de formas alternas y de consistencia interna se encuentran

en general en el rango de .70 para las formas individuales de las Escalas 2 y 3. Las confiabilidades de la prueba completa son más altas, en general a la mitad del rango de .80. Estos resultados se basan en docenas de estudios con miles de personas e indican un grado respetable de confiabilidad para un instrumento tan corto (IPAT, 1973).

La validez de la CFIT como medida de la inteligencia general se ha establecido más allá de cualquier duda razonable. Las puntuaciones de la CFIT se correlacionan a la mitad del rango de .80 con el factor general de inteligencia y muestran relaciones consistentemente fuertes —en gran medida en los rangos de .70 y .80— con otras medidas convencionales de inteligencia (WAIS, WISC, Matrices Progresivas de Raven, Stanford-Binet, Otis y la Batería de Prueba de Aptitud General; véase IPAT, 1973, p.11). No hay duda de que la CFTI es una prueba bien diseñada, útil y válida de la inteligencia.

Pero, ¿la CFIT es una **prueba culturalmente justa**, como afirma su nombre? Un objetivo de este instrumento era el de "reducir al mínimo las influencias irrelevantes del aprendizaje cultural y del clima social" y, por tanto, producir una "separación más evidente entre la capacidad natural y el aprendizaje específico" (IPAT, 1973). Por desgracia, la evidencia disponible indica que la CFIT no es más exitosa que las medidas tradicionales en la búsqueda de un método justo culturalmente para la medición de la inteligencia (Koch, 1984). Por ejemplo, Willard (1968) encontró que 83 niños afroestadounidenses con desventajas culturales obtuvieron puntuaciones aproximadamente iguales en la Stanford-Binet ($X = 68.1$) y en la CFIT ($X = 70.0$). Lo que es más, 14 de los niños llegaron al nivel inferior de la CFIT y recibieron la puntuación más baja de la prueba con un CI de 57, mientras las puntuaciones de la Stanford-Binet se distribuyeron en un patrón más similar a una curva con forma de campana.

COMENTARIO ACERCA DE LA CFIT

La CFIT es una excelente medida breve, no verbal, de la inteligencia general. Incluso cuando se emplean la Forma A y la Forma B para obtener lo que se conoce como prueba completa, la CFIT puede aplicarse a grupos grandes en menos de una hora. Una advertencia importante para los usuarios de la prueba es que la CFIT no ha logrado el encomiable objetivo de producir una prueba culturalmente justa. Lo que es más, el objetivo en sí puede ser quimérico:

> . . .las culturas difieren con respecto a la importancia que le dan a la competencia con los compañeros en la ejecución de tareas o en la solución de problemas, a la velocidad o calidad de la ejecución y a una variedad de otras conductas relacionadas con pruebas. Algunas culturas enfatizan la solución concreta de los problemas más que la abstracta, llegando con frecuencia al grado de que un problema carece de significado excepto en un entorno concreto. El mero concepto de contestar una prueba inventada de manera artificial carece de sentido en dichas situaciones (Koch, 1984).

Existen dudas de que incluso sea posible la existencia de una prueba culturalmente justa. En ediciones futuras, los creadores de la CFIT harían bien en cambiarle el nombre a su prueba, de modo que los usuarios poco experimentados no le confieran propiedades imaginarias a este instrumento.

Aunque la CFIT es una prueba valiosa, es necesaria una revisión y el establecimiento de nuevas normas. Tiene una apariencia más bien anticuada. Algunos de los dibujos para los reactivos de la prueba son tan pequeños que sólo las personas con visión perfecta pueden inferir las relaciones figurativas representadas en los componentes de los reactivos. Las anteriores muestras de estandarización se han especificado de manera deficiente y parecerían ser muestras de conveniencia más que representaciones estratificadas cuidadosamente seleccionadas de la población en conjunto.

MATRICES PROGRESIVAS DE RAVEN (RPM)

Presentadas originalmente en 1938, las Matrices Progresivas de Raven (RPM) son una

prueba no verbal del razonamiento inductivo que se basa en estímulos de figuras (Raven, Court y Raven, 1986, 1992). Esta prueba ha sido muy popular en investigación básica y también se le utiliza en ambientes institucionales con propósitos de detección intelectual.

ANTECEDENTES Y DESCRIPCIÓN

En un principio, la prueba RPM se diseñó como una medida del factor *g* de Spearman (Raven, 1938). Por esta razón, Raven eligió un formato especial de prueba que supuestamente requería el ejercicio de *g*. Se recuerda al lector que Spearman definió a *g* como la "deducción de correlatos". El término **deducción** se refiere al proceso de inferencia de relaciones con base en semejanzas fundamentales percibidas entre estímulos. En particular, para responder de manera correcta a las RPM, los individuos examinados deben identificar un patrón recurrente de relación entre los estímulos figurativos organizados en una matriz de 3 x 3. Los reactivos se clasifican en orden de dificultad creciente, de aquí la referencia a las matrices progresivas.

La prueba de Raven es en realidad una serie de tres instrumentos diferentes. Gran parte de la confusión acerca de la validez, la estructura factorial y similares se deriva de la suposición injustificada de que las tres formas deberían producir datos equivalentes. Se alienta al lector a abandonar estas hipótesis sin fundamento. Aunque las tres formas de la RPM se parecen unas a otras, pueden existir diferencias sutiles en las estrategias de solución de problemas que requieren cada una de ellas.

Matrices Progresivas en Color es una prueba con 36 reactivos diseñada para niños de 5 a 11 años de edad. Raven incorporó colores en su versión de la prueba para ayudar a mantener la atención de los sujetos jóvenes. Las normas de Matrices Progresivas Estándar se establecieron para niños de seis años en adelante, aunque la mayor parte de los reactivos son tan difíciles que la prueba es más adecuada para adultos. Esta prueba consta de 60 reactivos agrupados en cinco conjuntos de 12 progresiones. Matrices Progresivas Avanzadas es similar a la versión estándar, pero tiene un nivel superior más alto. La versión avanzada está formada por 12 problemas en el Conjunto I y 36 problemas en el Conjunto II. Esta forma es especialmente adecuada para individuos con un intelecto superior.

CARACTERÍSTICAS TÉCNICAS

En Raven y Summers (1986) se presentan las normas para Matrices Progresivas en Color y Estándar, obtenidas con grandes muestras en EUA. Se incluyen normas independientes para mexicano-estadounidenses y afroestadounidenses. Aunque no se hizo el intento de utilizar un procedimiento de muestreo aleatorio estratificado, la selección de distritos escolares fue tan ampliamente variada que las normas estadounidenses para niños parecen sólidas en un nivel razonable. Sattler (1988) resume las normas pertinentes para todas las versiones de la RPM. Recientemente, Raven, Court y Raven (1992) produjeron nuevas normas para la prueba de Matrices Progresivas Estándar, pero Gudjonsson (1995) ha hecho surgir la duda de que estos datos se hayan visto comprometidos debido a que no hubo vigilancia durante la prueba.

En cuanto a Matrices Progresivas en Color, se informan confiabilidades de división en mitades dentro del rango de .65 a .94, donde los individuos más jóvenes producen valores más bajos (Raven, Court y Raven, 1986). Para Matrices Progresivas Estándar, un valor típico de división en mitades es .86, aunque en personas más jóvenes se encuentran valores más bajos (Raven, Court y Raven, 1983). Las confiabilidades *test-retest* para las tres formas varían de manera considerable de una muestra a otra (Burke, 1958; Raven, 1965; Raven *et al.*, 1986). En el caso de adultos normales en los últimos años de la adolescencia y mayores, son típicos los coeficientes de confiabilidad entre .80 y .93. Sin embargo, para niños preadolescentes, se informan coeficientes de confiabilidad tan bajos como .71. Así, cuando se trata de individuos más jóvenes, es posible que la RPM no posea

la suficiente confiabilidad como para justificar su utilización en la toma de decisiones sobre individuos.

Los estudios analítico-factoriales de la RPM proporcionan poco o ningún apoyo a la intención original de la prueba de medir un constructo unitario (factor *g* de Spearman). Varios estudios sobre Matrices Progresivas en Color revelan tres factores ortogonales (Carlson y Jensen, 1980; Wiedl y Carlson, 1976). El Factor I incluye en gran medida, los reactivos muy difíciles y se le puede denominar cierre y razonamiento abstracto por analogías. El Factor II se cataloga como integración de patrones a través de identidad y cierre. El Factor III consta de los reactivos más fáciles y se define como una simple integración de patrones (Carlson y Jensen, 1980). En resumen, los reactivos muy fáciles y muy difíciles en Matrices Progresivas en Color parecen vincularse con procesos intelectuales diferentes.

Matrices progresivas avanzadas se descompone en dos factores que pueden tener validez predictiva por separado (Dillon, Pohlmann y Lohmann, 1981). El primer factor comprende reactivos en los que la solución se obtiene a través de la suma o resta de patrones (figura 6–10a). Los individuos que tienen un buen desempeño en estos reactivos pueden sobresalir en cuanto a la toma de decisiones rápidas y en situaciones donde deben percibirse las relaciones parte-todo. El segundo factor se compone de reactivos en los que la solución se basa en la capacidad para percibir la progresión de un patrón (figura 6–10b). Las personas que tienen un desempeño apropiado en estos reactivos pueden tener buena capacidad mecánica, así como habilidades adecuadas para estimar el movimiento proyectado y realizar rotaciones mentales. Sin embargo, las habilidades representadas por cada factor son, por el momento, meras conjeturas y requieren confirmación independiente.

Un enorme conjunto de investigaciones publicadas tiene que ver con la validez de la RPM. Burke (1958) proporciona un buen resumen de los primeros datos, y las conclusiones más recientes pueden encontrarse recopiladas en Court (1982) y en los manuales actuales de

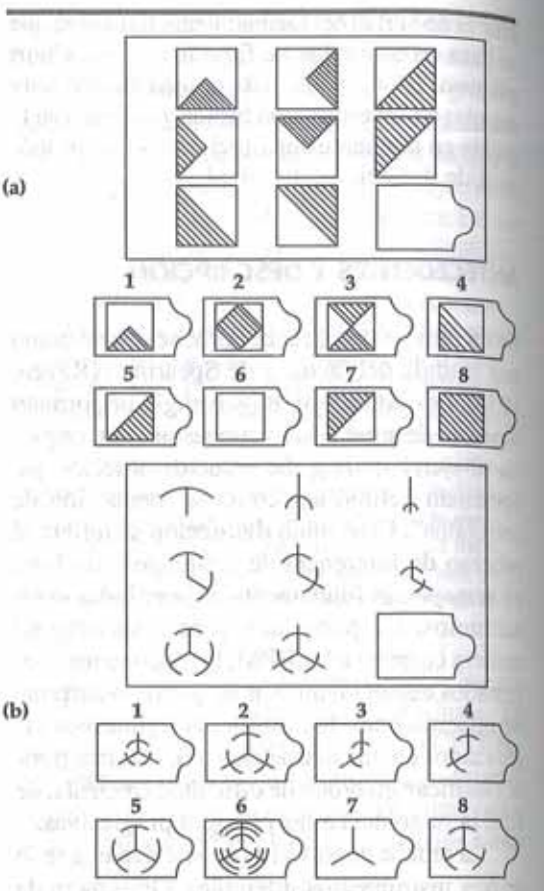


Figura 6–10. Matrices Progresivas de Raven: reactivos típicos.

la RPM (Raven y Summers, 1986; Raven, Court y Raven, 1983, 1986, 1992). En general, los coeficientes de validez en relación con pruebas de aprovechamiento van del rango de .30 al de .60. Como podría esperarse, estos valores son un poco más bajos que los encontrados en las pruebas más tradicionales (con cargas verbales) de la inteligencia. Los coeficientes de validez con otras pruebas de inteligencia van del rango de .50 al de .80. También, como podría esperarse, las correlaciones tienden a ser más altas con las pruebas de ejecución que con las pruebas verbales. En un enorme estudio en el que participaron miles de escolares, Saccuzzo y Johnson (1995) concluyeron que Matrices Progresivas Estándar y el WISC-R mostraron una validez predictiva aproximadamente igual y

que no existe evidencia de validez diferencial entre los ocho grupos étnicos diferentes.

Johnson, Saccuzzo y Guertin (1994) lograron lo que (casi) es imposible al desarrollar una forma alterna verdaderamente comparable de Matrices Progresivas Estándar. Para cada uno de los 60 reactivos originales, desarrollaron un reactivo similar que era comparable en términos de nivel de dificultad y estrategia cognitiva subyacente requerida para la solución. Un análisis de confiabilidad de formas alternas con un grupo diverso de 449 niños que tomaron ambas pruebas en orden contrabalanceado reveló un coeficiente de confiabilidad de .90, que está al nivel de los datos de *test-retest* inmediato. En esta misma muestra, la distribución de puntuaciones no mostró diferencias en cuanto a desviación estándar, asimetría y rango de dificultad de los reactivos. La media de respuestas correctas fue de 36.1 en MPE y de 35.5 en la nueva prueba. En resumen, las dos versiones de la prueba son casi idénticas en características psicométricas generales y también en nivel de dificultad. La nueva prueba promete satisfacer un importante papel en los estudios de investigación que requieren de la repetición de prueba.

COMENTARIO ACERCA DE LA RPM

Aunque la RPM no ha cumplido las expectativas originales de medir el factor *g* de Spearman, la prueba es, sin embargo, un índice útil del razonamiento no verbal, figurativo. La actualización reciente de las normas fue un avance bien acogido para esta prueba tan conocida, en el sentido de que muchos usuarios estadounidenses estaban recelosos de las normas británicas por considerarlas anticuadas y limitadas. Sin embargo, las normas adultas de Matrices Progresivas Estándar y Avanzadas continúan siendo bastante restringidas.

La RPM es particularmente valiosa como prueba complementaria para niños y adultos con discapacidades auditivas, del lenguaje y físicas. Con frecuencia, estos individuos presentan dificultades cuando se les evalúa con medidas tradicionales que requieren atención auditiva, expresión verbal o manipulación física. En contraste, la RPM puede explicarse a través de pantomima, si ello es necesario. Lo que es más, la única vía expresiva que se requiere del examinando es una marca con un lápiz o un ademán que denote la alternativa elegida. Por tales razones, la RPM es ideal para examinar a personas que tienen un dominio limitado del idioma inglés. De hecho, la RPM está tan restringida en términos culturales como es posible: el protocolo de prueba no contiene una sola palabra en algún idioma.

Mills y Tissot (1995) encontraron que Matrices Progresivas Avanzadas identificaba como dotados a una mayor proporción de niños minoritarios que una medida más tradicional de la aptitud académica (la Prueba de Capacidad Escolar y Universitaria).

Finalmente, debe expresarse una advertencia: algunos individuos sumamente inteligentes y de alto funcionamiento tienen un desempeño muy inferior en la RPM. Gregory y Gernert (1990) sometieron a prueba (con una variante de la RPM) a casi 100 miembros de cuerpos docentes universitarios. Un investigador sobresaliente y que había llegado a nivel vicepresidencial, no tuvo la menor idea de cómo resolver los problemas de la RPM y obtuvo puntuaciones a un nivel aleatorio. Algunas personas de inteligencia superior al promedio simplemente no tienen un buen desempeño en tareas de razonamiento figurativo. Los examinadores harían bien en cuestionar la validez de una baja puntuación en el caso de un individuo que en otros sentidos es muy exitoso.

PERSPECTIVA SOBRE LAS PRUEBAS CULTURALMENTE JUSTAS

La *Culture Fair Intelligence Test* (CFIT; Prueba de Inteligencia Culturalmente Justa de Cattell) y las Matrices Progresivas de Raven (RPM) se citan con frecuencia como ejemplos de pruebas culturalmente justas, lo cual es un concepto con una historia larga y confusa. Aquí se intentará aclarar algunos términos y problemas.

El primer asunto a analizar es que las pruebas de inteligencia son tan sólo muestras de aquello que las personas saben y pueden hacer. No se deben materializar y sobrevalorar las pruebas de inteligencia. Las pruebas nunca son muestras de inteligencia innata o de conocimiento independiente de la cultura. Todo conocimiento se basa en la cultura y se adquiere a través del tiempo. Como señala Scarr (1994), no existe nada que se asemeje a una prueba culturalmente libre.

Pero, ¿qué sucede con una prueba culturalmente justa que plantea problemas que son igualmente familiares (o no familiares) para todas las culturas? Ésta parecería ser una posibilidad más realista que una prueba culturalmente independiente, pero incluso en este caso los escépticos pueden tener dudas. Considérese la pregunta de qué es lo que **significa** una prueba, lo cual difiere de una cultura a otra. En teoría, una prueba de matrices parecería igualmente justa para la mayoría de las culturas. Pero en la práctica surgen problemas sobre equidad. Las personas criadas en culturas occidentales están entrenadas en un pensamiento lineal, convergente. Se sabe que el propósito de una prueba es encontrar una sola respuesta adecuada y hacerlo con rapidez. Si se examina la matriz de 3 × 3 de izquierda a derecha y de arriba hacia abajo, buscando los principios lógicos invocados en la sucesión de formas, ¿puede suponerse que las personas criadas en Nepal o en Nueva Guinea o incluso en las partes más remotas de Idaho harán lo mismo? La prueba puede significar algo diferente para ellos. Quizás la aborden como una medida de la progresión estética, más que como una sucesión lógica. Quizá la consideren como una bobería de tal tipo que no es digna de un esfuerzo intelectual intenso. Suponer que una prueba es igualmente justa para todos los grupos culturales tan sólo porque los estímulos son igualmente familiares (o no familiares), es algo inapropiado. Se puede hablar acerca de grados de justicia (o injusticia) cultural, pero el concepto de que cualquier prueba sea culturalmente justa en términos absolutos es, con toda seguridad, un equívoco.

RESUMEN

1. Las pruebas grupales de inteligencia difieren de las individuales en cinco aspectos: formato de opción múltiple contra formato de preguntas abiertas, calificación mecánica objetiva contra calificación del examinador, aplicación grupal contra individual, empleo con fines de detección contra empleo para planificación de regularización y muestras enormes de estandarización contra muestras simplemente grandes.
2. La ventaja obvia de las pruebas grupales es que puede examinarse con rapidez y eficacia a grandes números de personas; la desventaja es que el individuo evaluado puede tener puntuaciones muy por debajo de su verdadera capacidad debido a problemas motivacionales o dificultad para seguir instrucciones.
3. La MAB es una prueba grupal de inteligencia, con preguntas de opción múltiple, diseñada como un equivalente de lápiz y papel para el WAIS-R. Los CI de la MAB y del WAIS-R se correlacionan en .82, .65 y .91 para las Escalas Verbal, de Ejecución y Total, respectivamente. La confiabilidad *test-retest* del instrumento es excelente y los análisis factoriales sustentan su validez de constructo.
4. La escala del *Shipley Institute of Living* (SILS) se propuso originalmente como un índice del deterioro intelectual. La SILS consta de una sección de vocabulario con 40 reactivos de opción múltiple y una sección de abstracciones con 20 reactivos de llenado de espacios en blanco. La prueba no ha funcionado bien como índice de organicidad, pero sí satisface una necesidad

de un instrumento breve de detección para la inteligencia general.

5. La *Cognitive Abilities Test* (CogAT) es representativa de muchas baterías escolares multinivel utilizadas en la actualidad. Las nueve subpruebas de la CogAT incluyen una batería Verbal, una batería Cognitiva y una batería No verbal. La prueba se normó de manera conjunta con dos pruebas de aprovechamiento, las pruebas Iowa de habilidades básicas y las de ejecución y aprovechamiento.
6. La *Culture Fair Intelligence Test* (CFIT) es una medida no verbal de la inteligencia fluida que intenta reducir al mínimo el prejuicio cultural. La CFIT es adecuada para individuos desde los cuatro años de edad hasta la edad adulta, y se presenta en tres versiones, cada una de las cuales consta de dos formas equivalentes; cada una incluye cuatro subpruebas: Series, Clasificación, Matrices y Condiciones.
7. La confiabilidad de la CFIT es excelente y sus puntuaciones se correlacionan en gran medida con otras pruebas de inteligencia muy respetadas. La CFIT es una buena medida de la inteligencia, pero es probable que tenga tantos límites culturales como las pruebas más tradicionales. La prueba necesita revisión y reestandarización.
8. Matrices Progresivas de Raven (RPM) se diseñó originalmente como una medida del factor *g* de Spearman y es una prueba no verbal del razonamiento inductivo que se basa en estímulos figurativos y que se presenta en tres versiones diferentes: Matrices Progresivas en Color (5 a 11 años de edad), Matrices Progresivas Estándar (seis años hasta la edad adulta) y Matrices Progresivas Avanzadas.
9. Aunque la RPM es un índice confiable y válido de la inteligencia, la prueba ha recibido poco apoyo como medida unitaria del factor *g*. En general, los análisis factoriales revelan 2 o 3 factores, incluyendo razonamiento por analogía e integración simple de patrones. La RPM es útil como prueba complementaria para personas con discapacidades auditivas, del lenguaje o físicas.
10. Las pruebas culturalmente justas son una abstracción idealizada imposible de lograr dentro del mundo real. Incluso el significado de una prueba puede diferir entre grupos culturales, lo cual afectará la validez de las comparaciones. Algunas pruebas son más justas en términos culturales que otras, pero no es posible que ninguna prueba sea igualmente justa para todos los grupos culturales.

TÉRMINOS Y CONCEPTOS CLAVE

Índice de deterioro intelectual p. 249
Prueba culturalmente justa p. 256

CAPÍTULO 7

Sesgo de una prueba y examen de poblaciones

TEMA 7A
Pruebas con poblaciones especiales

Orígenes de las pruebas para poblaciones especiales
Obligación legal de evaluar a las personas con discapacidades
Pruebas que no requieren del lenguaje
Pruebas que no requieren la lectura y pruebas con reducción motora
Ejemplo de caso 7–1. El reto de la evaluación en la parálisis cerebral
Pruebas para individuos con alteraciones visuales
Evaluación de la conducta adaptativa en el retraso mental
Resumen
Términos y conceptos clave

Las pruebas individuales y grupales de la inteligencia que se revisaron en capítulos anteriores son adecuadas para personas con capacidades normales o casi normales en habla, audición, visión, movimiento y capacidad intelectual general. No obstante, no todo individuo examinado cae dentro del espectro habitual de capacidades físicas y mentales. En razón a su juventud, discapacidad física, intelecto disminuido o desventajas de lenguaje, una gran proporción de la población queda fuera del alcance de las pruebas y procedimientos tradicionales. Según el *Census Bureau* de EUA, cerca de 25 millones de estadounidenses (1 de cada 10) tiene una discapacidad **grave** que le impide realizar una o más actividades o funciones (www.census.gob, 1998). Este estimado no incluye a personas que viven en instituciones. En estos casos especiales se necesitan pruebas originales para la evaluación válida. En el tema 7A, Pruebas con poblaciones especiales, se analizan los instrumentos diseñados para las consultas

excepcionales y difíciles, como las de personas con discapacidades sensoriales/motoras, inmigrantes recientes que provienen de países que no son de habla inglesa, e individuos con deficiencias intelectuales importantes. En el Tema 7B, Predisposición de prueba y otras controversias, se continúa con un tema que requiere cautela y en el que se presentan varias preocupaciones acerca del uso y significado de las calificaciones en pruebas de inteligencia.

ORÍGENES DE LAS PRUEBAS PARA POBLACIONES ESPECIALES

En el decenio de 1950, surgió en EUA un compromiso renovado hacia las necesidades y derechos de las personas discapacitadas física y mentalmente (Maloney y Ward, 1979; Patton, Payne y Beirne-Smith, 1986). Las actitudes de la sociedad hacia aquellos que tienen necesidades especiales cambiaron de un menosprecio evidente a una postura de mayor apoyo, lo que favoreció nuevos programas e iniciativas en favor de las personas con discapacidad. El progreso ha sido lento, pero ya no sorprende ver instalaciones sanitarias con acceso para sillas de ruedas para personas con impedimentos físicos, libros impresos en tipo más grande para individuos con alteraciones visuales o programas de televisión con subtítulos en el mismo idioma para personas con discapacidades auditivas. Lo que es más, cada vez se satisfacen más las necesidades especiales de los ciudadanos con retraso mental a través de pequeños centros de atención comunitaria, en vez de instituciones masivas e impersonales.

A principios del decenio de 1970, la renovada preocupación por las necesidades de las personas con discapacidad se tradujo en una legislación federal. En 1973 se aprobó la **Ley pública 93-112**, que sirvió como una "Declaración de derechos" para los individuos con discapacidades. Esta ley prohibió la discriminación con base en la discapacidad. Dos años después, se aprobó el Decreto de educación para todos los niños con discapacidades (Ley pública 94-142), lo cual constituyó un hito. Esta ley decretaba que todos los niños discapacitados en edad escolar recibieran evaluación y oportunidades educativas apropiadas. En particular, se instruyó a los psicólogos para que evaluaran a los niños en todas las áreas de posible limitación —mental, conductual y física— y que utilizaran instrumentos validados para dichos propósitos expresos.

En este tema se analizan las pruebas que pueden utilizarse para la evaluación de personas con discapacidades sensoriales, motoras o mentales. No obstante, antes de analizar las pruebas específicas, se revisarán ciertas distinciones entre los tipos de pruebas disponibles para las evaluaciones excepcionales. El lector también apreciará un resumen de los ordenamientos legales que han dado forma a las prácticas de evaluación con individuos discapacitados.

ENFOQUES PARA LA EVALUACIÓN DE POBLACIONES ESPECIALES

Las pruebas especiales se diseñaron inicialmente a principios del siglo XX para examinar a inmigrantes que no hablaban inglés, personas con sordera e individuos con defectos del habla (DuBois, 1970). Estos primeros instrumentos especiales eran, en gran medida, pruebas de ejecución o pruebas que no utilizaban el lenguaje y podían aplicarse a través de mímica. La persona examinada manipulaba objetos o utilizaba lápiz y papel para realizar tareas de fácil comprensión, como trazar un camino a través de un laberinto.

También se han diseñado instrumentos especiales para individuos que no pueden leer, pero poseen la capacidad para comprender el inglés hablado. Estas pruebas que no requieren de la lectura se dirigen a niños pequeños y a otras personas iletradas que, sin embargo, pueden comprender y seguir instrucciones verbales. Muchas pruebas que no requieren la lectura, implican la manipulación de objetos. No obstante, una prueba de este tipo también puede evaluar las habilidades de comprensión del lenguaje al utilizar un formato de vocabulario

con imágenes: el examinador dice una palabra y la persona evaluada señala un dibujo que representa la palabra entre un conjunto de imágenes. Más adelante se analizan varias pruebas de vocabulario con ilustraciones.

Una prueba con reducción motora requiere el mínimo de salida motriz para la respuesta. En una prueba de este tipo, la persona sólo tiene que señalar o hacer un ademán hacia la respuesta correcta entre varias alternativas. Por ejemplo, un individuo con parálisis cerebral podría responder a los reactivos de vocabulario con imágenes al colocar una mano sobre la alternativa elegida. Algunas pruebas que no requieren lectura —en particular aquellas que utilizan un formato de vocabulario con imágenes— también son con reducción motora.

Por último, se debe mencionar que varios dispositivos importantes de evaluación en realidad no son pruebas. Un inventario del desarrollo es un instrumento estandarizado para observar y valorar el desarrollo conductual en lactantes y niños pequeños. En general, estos instrumentos analizan los principales hitos del desarrollo, como sentarse por sí solo, ponerse de pie sin ayuda, etc. Una característica de dichas herramientas es que la persona "examinada" no realiza la prueba por sí misma o, en todo caso, hace algo fuera de lo común. Un inventario del desarrollo es, en realidad, sólo una forma estructurada de observación. De la misma manera, una escala conductual es un instrumento para determinar el perfil de habilidades (y quizás excesos) conductuales exhibidos por un niño o un adulto con retraso. En general, las escalas conductuales las llena un adulto familiarizado con la persona (padre, maestro o psicólogo).

OBLIGACIÓN LEGAL DE EVALUAR A LAS PERSONAS CON DISCAPACIDADES

Muchas de las prácticas de evaluación con personas que tienen discapacidades son resultado directo de la legislación y de casos judiciales. Como antecedente para el análisis de las pruebas y procedimientos específicos, se ofrece una breve revisión de las leyes públicas referentes a la evaluación de personas con discapacidades. La cobertura es propositivamente breve. Los lectores pueden encontrar análisis más amplios en Bruyere y O'keeffe (1994), Salvia e Ysseldyke (1991), y Sattler (1988) .

LEY PÚBLICA 94-142

En 1975, el Congreso de EUA aprobó una ley obligatoria de educación especial, la **Ley Pública 94-142**, conocida como Decreto de educación para todos los niños con discapacidades.* Según Ballard y Zettel (1977) esta ley se diseñó para satisfacer cuatro metas principales:

1. Asegurar que los servicios de educación especial estén disponibles para todos los niños que los requieran.
2. Garantizar que las decisiones sobre los servicios para estudiantes discapacitados sean justas y apropiadas.
3. Establecer requisitos específicos de administración y auditoría para la educación especial.
4. Proporcionar fondos federales para ayudar a los estados a educar a los estudiantes con discapacidad.

Muchas de las prácticas dentro de la evaluación de las personas con impedimentos se derivan directamente de las disposiciones de la Ley Pública 94-142. Por ejemplo, la ley especifica que cada estudiante con discapacidad debe recibir un Plan Educativo Individualizado (IEP, del inglés *individualized education plan*) con base en una evaluación completa de parte de un equipo multidisciplinario. El IEP debe delimitar los objetivos a largo y corto plazo y especificar los planes para lograrlos; además, indicar la manera en que se evaluará el progreso hacia estos objetivos. Los padres deben participar de

* Cada ley del Congreso recibe dos números, uno que se refiere a la legislatura particular que la aprobó, y el otro a la ley en sí; por tanto, la Ley Pública 94-142 es la ley 142 aprobada por la Legislatura número 94.

manera íntima en este proceso y aprobar los aspectos particulares del IEP.

En relación con estas prácticas de prueba, la LP 94-142 incluye varias disposiciones diseñadas para asegurar que los procedimientos y actividades de evaluación sean justos, equitativos y no discriminatorios. Salvia e Ysseldyke (1988) resumen estas medidas como sigue:

1. Las pruebas deben seleccionarse y aplicarse de tal modo que no sean discriminatorias en términos raciales y culturales.
2. Hasta donde sea posible, debe evaluarse a los estudiantes en su lengua materna o forma de comunicación
3. Las pruebas deben validarse para el propósito específico para el cual se utilizan.
4. La aplicación de las pruebas debe llevarla a cabo personal entrenado, de conformidad con las instrucciones que proporciona el autor de la prueba.
5. Las pruebas utilizadas con estudiantes deben incluir aquéllas diseñadas para proporcionar información sobre necesidades educativas específicas y no sólo un cociente de inteligencia en general.
6. Las decisiones sobre estudiantes deben basarse en algo más que el desempeño en una sola prueba.
7. Las valoraciones debe realizarlas un equipo multidisciplinario que incluya al menos a un maestro u otro especialista con conocimientos en el área donde se sospecha la existencia de discapacidad.
8. La evaluación de los niños debe realizarse en todas las áreas relacionadas con una discapacidad específica, incluyendo —cuando sea apropiado— salud, visión, audición, estado social y emocional, inteligencia general, desempeño académico, habilidades de comunicación y habilidades motoras.

La LP 94-142 también contiene una disposición para que se coloque a los estudiantes con discapacidad en el ambiente menos restrictivo —aquel que permita la mayor oportunidad posible de interacción con estudiantes no discapacitados. La instrucción por separado ha de darse sólo cuando la naturaleza o gravedad del impedimento es tal que no puedan lograrse las metas educativas en el salón de clases regular. Por último, la ley contiene una cláusula que garantiza una audiencia imparcial para resolver conflictos entre los padres de niños con discapacidad y el sistema escolar.

En general, las medidas de la LP 94-142 han proporcionado un fuerte ímpetu al desarrollo de pruebas especializadas con diseño, normas y validez establecidos para niños con discapacidades específicas. Por ejemplo, en la evaluación de un niño con alteraciones visuales, las disposiciones de la LP 94-142 virtualmente ordenan que el examinador utilice una prueba con normas adecuadas, diseñada justo para esa población, en vez de basarse en instrumentos tradicionales.

LEY PÚBLICA 99-457

En 1986, el Congreso de EUA aprobó varias enmiendas al Decreto de educación para todos los niños con discapacidades, aumentando las disposiciones de la LP 94-142 para incluir a los niños preescolares con discapacidades. La **Ley Pública 99-457** requiere que los estados proporcionen educación pública gratuita y apropiada a niños discapacitados entre los 3 y 5 años de edad; también ordena subsidios para los estados que ofrecen servicios educativos interdisciplinarios a lactantes e infantes discapacitados y sus familias, con lo cual se establece un incentivo para los estados que proporcionan atención a niños con discapacidades, desde el nacimiento hasta los dos años. Esta ley también impulsa el desarrollo y validación de pruebas e inventarios conductuales para lactantes. Después de todo, la identificación precisa y temprana de los niños en riesgo parecería ser un primer paso importante en la intervención interdisciplinaria eficaz.

DECRETO SOBRE LOS ESTADOUNIDENSES CON DISCAPACIDADES

El **Decreto sobre los estadounidenses con discapacidades** (ADA, del inglés *Americans*

with Disabilities Act), aprobado en 1990, prohíbe la discriminación, tanto en el sector público (p. ej., instancias de gobierno y entidades que reciben subsidios federales) como en el privado (p. ej., corporaciones y otros empleadores que obtienen lucro), en contra de individuos calificados con discapacidades. Una disposición del ADA en cuanto a las pruebas psicológicas, es que las instituciones deben realizar arreglos razonables para evaluación de las personas con discapacidades, que incluyen cambios en el medio de la prueba, límites de tiempo e incluso (en pocas ocasiones) en el contenido de la misma. Un ejemplo de un arreglo apropiado en el medio de la prueba es la presentación en audiocinta de los reactivos para personas con alteraciones visuales. Por otro lado, cambiar una prueba de su versión impresa a la de lenguaje de signos para personas con deterioro auditivo se consideraría una traducción a otro idioma y no un simple cambio de medio.

En la mayor parte de los arreglos de prueba que exige el ADA, es necesario cambiar los límites de tiempo aumentándolo. Esto hace surgir problemas en la interpretación de la prueba, en especial cuando es esencial un estricto límite de tiempo para la validez de una prueba. Por ejemplo, Willingham, Ragosta, Bennett y colaboradores (1988) encontraron que el aumento en límites de tiempo en la SAT reducía de manera significativa la validez de la prueba como medio de predicción de las calificaciones en el primer año de universidad. Esto fue particularmente cierto en el caso de personas evaluadas que tenían discapacidades para el aprendizaje, cuyas calificaciones en el primer año fueron mucho menores a las pronosticadas por sus calificaciones en la SAT. Así, aunque parece justo proporcionar tiempo adicional en una prueba cuando se ha cambiado el medio de la misma (p. ej., preguntas grabadas en audiocinta que reemplazan la versión impresa), desde un punto de vista psicométrico, el reto consiste en determinar **qué tanto** tiempo adicional debería proporcionarse de modo que la prueba modificada sea comparable con la versión original. Nester (1994) y Phillips (1994) proporcionan perspectivas cuidadosas acerca del rango de arreglos razonables requeridos por el ADA.

Las implicaciones del ADA con respecto a las pruebas de personalidad son poco claras, pero pueden ser sustanciales. Una medida establecida por éste es que no pueden aplicarse pruebas médicas antes de ofrecer un empleo. Si las pruebas de personalidad se interpretan como de naturaleza "médica", esto podría reducir de manera abrupta el uso de las mismas en algunos exámenes relacionados con el empleo de personas con discapacidades (Klimoski y Palmer, 1994). Al momento de la publicación de la edición en inglés de este libro, la interpretación del ADA todavía sufre modificaciones por los desafíos legales y pautas administrativas.

Ahora que se han resumido los antecedentes legales de la evaluación de personas con necesidades especiales, la atención se centra en una revisión de los instrumentos típicos utilizados para examinar a individuos con discapacidades. Se organizó la revisión alrededor de los siguientes temas: pruebas que no requieren el lenguaje ni la lectura, pruebas con reducción motora, pruebas para individuos con alteraciones visuales y evaluación de la conducta adaptativa en el retraso mental.

PRUEBAS QUE NO REQUIEREN DEL LENGUAJE

Como recordará el lector, las pruebas que no requieren lenguaje demandan poco o ningún lenguaje escrito o hablado de parte del examinador o de la persona evaluada. Así, son particularmente adecuadas para la evaluación de personas que no hablan el inglés, individuos con discapacidades del habla y personas con pocas habilidades de lenguaje. Estos instrumentos también pueden utilizarse con personas que no tienen discapacidades.

ESCALA INTERNACIONAL LEITER DE EJECUCIÓN-REVISADA (LIPS-R)

La *Leiter International Performance Scale-Revised* (Escala Internacional Leiter de Ejecución-

Revisada; LIPS-R; Roid y Miller, 1997) es una revisión reciente de una prueba típica y muy respetada de la inteligencia no verbal y de las capacidades cognitivas (Leiter, 1948, 1979). Leiter diseñó una edición experimental de la prueba en 1929 para evaluar la inteligencia de personas con alteraciones auditivas, trastornos del habla, bilingües o que no hablaban inglés. La escala se probó en un programa piloto con varios grupos étnicos en Hawai, incluyendo a niños de ascendencia japonesa y china. La primera edición se basó en los resultados de prueba para niños estadounidenses, estudiantes de bachillerato y reclutas del ejército en la segunda Guerra Mundial.

Aunque muy elogiada y de amplio uso después de su publicación inicial, esta prueba recibió fuertes críticas en años recientes debido a la deficiencia de sus ilustraciones y a sus normas obsoletas. La revisión de la escala Leiter responde hábilmente a todas las críticas y la LIPS-R merece una amplia utilización como medida culturalmente reducida de la inteligencia no verbal.

Un aspecto notable de la escala Leiter es la completa eliminación de instrucciones verbales. La escala no requiere una sola palabra hablada por parte del examinador o del examinando. Con un rango de edades de los 2 a los 20 años y 11 meses, la Leiter-R es particularmente idónea para niños y adolescentes cuyas habilidades en el idioma inglés son pocas. Esto incluye a niños con cualesquiera de los siguientes aspectos: no angloparlantes, con autismo, lesión cerebral traumática, alteraciones del habla, problemas auditivos y que provienen de un ambiente económicamente bajo. La prueba también es útil para la evaluación de problemas de atención, como se describe a continuación.

La tarea de la prueba consiste en que el niño o el adolescente equipare pequeñas tarjetas laminadas bajo las ilustraciones correspondientes en un folleto colocado en forma de atril (figura 7–1). La prueba es sin límite de tiempo. Debido a que los reactivos iniciales son más que obvios, la mayoría de los individuos examinados comprenden con rapidez el propósito sin necesidad de una demostración con señas.

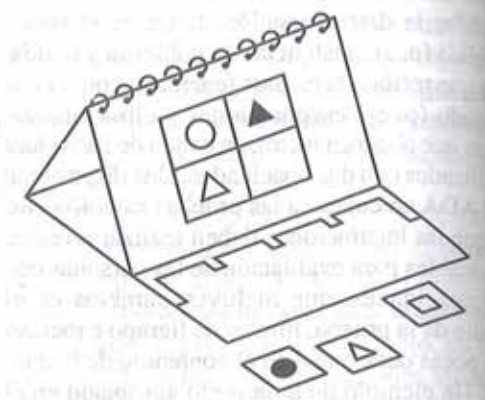


Figura 7–1. Reactivo característico de la Escala Internacional Leiter de Ejecución-Revisada.

La LIPS-R contiene 20 subpruebas organizadas en cuatro dominios: Razonamiento, Visualización, Memoria y Atención. No todas las subpruebas se aplican a todos los niños. Por ejemplo, la subprueba de Rotación de Figura es demasiado difícil para niños de dos años de edad, y la subprueba de Reconocimiento Inmediato es demasiado fácil para los adolescentes evaluados. Las cuatro subpruebas de Razonamiento incluyen clasificación y analogías de diseño. Las seis subpruebas de Visualización incluyen equiparación, figura-fondo, doblado de papel y rotación de figura. Las ocho subpruebas de Memoria incluyen capacidad de memoria, memoria espacial, memoria asociativa y memoria de reconocimiento demorado. Las dos subpruebas de Atención constan de una prueba de subrayado (p. ej., marcar todos los cuadrados impresos en una página completa de formas geométricas) y una medida de atención dividida (p. ej., observar una muestra en movimiento y clasificar simultáneamente tarjetas de manera correcta).

La LIPS-R produce un CI mixto con la media familiar de 100 y desviación estándar de 15. La prueba también produce puntuaciones escalares de subprueba con media de 10 y desviación estándar de 3, así como una variedad de puntuaciones mixtas útiles para el diagnóstico clínico. Las normas de la prueba se

establecieron con más de 2 000 niños y adolescentes, de los 2 a los 21 años de edad. Con los datos estadísticos del censo de 1993, se estratificó cuidadosamente a estos individuos según raza, edad, género, clase social y región geográfica. La confiabilidad de consistencia interna para las subpruebas, calificaciones de dominio y calificaciones de CI es excelente. Los coeficientes α típicos se encuentran desde la parte alta del nivel de .80 para las subpruebas hasta la parte baja del de .90 para las calificaciones de dominio y de CI. Extensos estudios sobre parcialidad de los reactivos revelan que éstos parecen funcionar de manera similar en grupos raciales diferentes (muestras de personas de raza blanca, afroestadounidenses e hispanos), es decir, no existe evidencia de parcialidad (definida como el funcionamiento diferencial de los reactivos). Junto con el hecho de que la prueba es totalmente no verbal, la ausencia de parcialidad de prueba indica que la escala LIPS-R es una buena elección para la evaluación culturalmente reducida de niños que provienen de minorías.

Aunque en este momento la investigación de correlación con esta escala es escasa, los estudios con la primera edición indican fuertes relaciones con las calificaciones de otras pruebas de inteligencia. Por ejemplo, la escala Leiter y el CI de Ejecución del WISC se correlacionan en un punto cercano al rango de .80; las correlaciones con el CI Verbal del WISC se localizan más en el rango de .60 (Arthur, 1950; Matey, 1984). Reeve, French y Hunter (1983) compararon la escala Leiter y la Stanford-Binet: Forma L-M como medios de predicción de las calificaciones en la Prueba de ejecución Metropolitan (MAT) de 60 alumnos de jardín de niños. Las correlaciones fueron de .77 entre la Stanford-Binet y el total de la MAT y de .61 entre la Leiter y el total de la MAT. Los autores señalan que aunque la Stanford-Binet resultó ser un medio de predicción marginalmente mejor de la ejecución individual, los niños con problemas auditivos, del habla o ambos pueden requerir la escala Leiter u otros instrumentos no verbales.

La LIPS-R es una revisión bien recibida de una prueba obsoleta. En manos de un clínico cuidadoso, es útil para la evaluación intelectual de niños con habilidades deficientes en el idioma inglés. Otros de sus usos incluyen la evaluación del trastorno por déficit de atención con hiperactividad (en este sentido son importantes las comparaciones de las subpruebas de Atención con otros dominios) y la valoración de la inteligencia dotada en niños pequeños (para esta aplicación resulta invaluable el límite superior extremadamente alto de la prueba). Mientras los revisores han advertido en contra de utilizar la escala Leiter original con propósitos de asignación o toma de decisiones (Sattler, 1988; Salvia e Ysseldike, 1991), la escala Leiter revisada está mejor en cuanto a calidad psicométrica y excelencia en la estandarización.

PRUEBAS DE DIBUJO DE LA FIGURA HUMANA

La mayoría de los niños disfrutan dibujando figuras humanas y lo hacen de manera rutinaria y espontánea. Desde principios del siglo XX, los psicólogos han intentado utilizar esta conducta casi instintiva como base para medir el desarrollo intelectual. La primera persona en utilizar el dibujo de la figura humana (DFH) como prueba estandarizada de inteligencia fue Florence Goodenough (1926). Su prueba, conocida como Test del Dibujo de la Figura Humana, fue revisada por Harris (1963) y se le cambió el título por el de Test de Dibujo Goodenough-Harris. Recientemente, Naglieri (1988) adaptó la técnica del DFH. El enfoque adicional de Gonzales (1986) no se revisa aquí. También debería mencionarse que los dibujos de la figura humana se utilizan **ampliamente** como medidas de adaptación emocional, pero aquí no se analizará esa aplicación.

El Test de Dibujo Goodenough-Harris es una prueba breve, no verbal, de inteligencia que puede aplicarse de manera individual o grupal. Goodenough (1926) publicó la primera edición de esta prueba, mientras Harris (1963) proporcionó importantes mejorías en cuanto a calificación y estandarización, incluyendo el uso de un CI de desviación. En términos estrictos, el

test Goodenough-Harris no satisface los criterios de una prueba que no requiere del lenguaje, en el sentido de que el examinador debe trasmitir ciertas instrucciones en español o por medio de un intérprete. No obstante, las instrucciones son breves y básicas ("quiero que hagas un dibujo de un hombre [o mujer]; haz el mejor dibujo que puedas"). El test Goodenough-Harris es, para propósitos prácticos, un instrumento que no requiere del lenguaje.

El propósito de esta prueba es el de medir la madurez intelectual, no la habilidad artística. Así, la guía de calificación enfatiza la precisión de la observación y el desarrollo del pensamiento conceptual. El niño recibe bonificación por incluir partes y detalles del cuerpo, así como por brindar perspectiva, proporción realista y dar la implicación de libertad de movimiento.

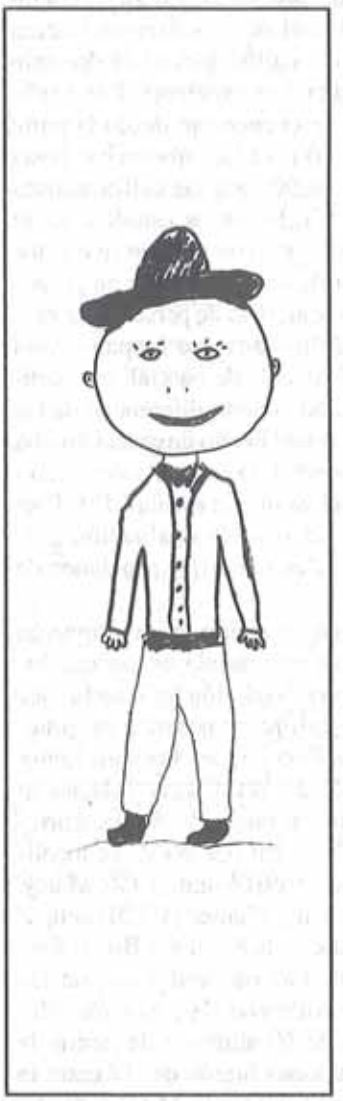
Los 73 elementos a calificar se seleccionaron según el siguiente criterio:

1. Los elementos deben mostrar un incremento regular y bastante rápido con la edad, en el porcentaje de niños que pasan el punto.
2. Los elementos deben mostrar una relación con alguna medida general de la inteligencia.
3. Los elementos deben diferenciar entre niños con altas calificaciones en la escala completa y aquellos que tienen bajas calificaciones en la misma (Harris, 1963).

En la figura 7–2 se presenta un dibujo de muestra con calificación reactivo por reactivo.

Además de la escala de Hombre, la revisión Harris (1963) también incluye dos formas adicionales: la escala de Mujer y la escala de Sí mismo. Para estas dos últimas escalas se instruye a las personas evaluadas a que hagan un dibujo de una mujer y un dibujo de sí mismos. Las calificaciones en las escalas de Hombre y de Mujer tienen una muy alta correlación para personas de cualquiera de ambos sexos ($r = .91$ a .98). Estas versiones pueden considerarse como formas equivalentes. La escala de Sí mismo tenía la intención de ser una prueba proyectiva del autoconcepto. No obstante, el

Haz un dibujo de un hombre. Haz el mejor dibujo que puedas. Asegúrate de dibujar el hombre completo, no sólo su cabeza y hombros.



1.	+	41.	0
2.	+	42.	0
3.	+	43.	0
4.	+	44.	+
5.	+	45.	+
6.	+	46.	+
7.	+	47.	+
8.	0	48.	0
9.	+	49.	0
10.	0	50.	0
11.	+	51.	0
12.	0	52.	0
13.	0	53.	0
14.	+	54.	+
15.	0	55.	+
16.	0	56.	+
17.	0	57.	0
18.	+	58.	+
19.	0	59.	0
20.	0	60.	0
21.	0	61.	0
22.	+	62.	+
23.	+	63.	+
24.	+	64.	+
25.	0	65.	0
26.	+	66.	0
27.	0	67.	+
28.	+	68.	+
29.	+	69.	0
30.	+	70.	0
31.	+	71.	0
32.	0	72.	0
33.	+	73.	0
34.	0		
35.	+		
36.	+		
37.	0		
38.	0		
39.	+		
40.	+		

Puntuación natural 37

Nota: Este esfuerzo de una niña de ocho años de edad se convierte a una Puntuación Estándar de 118 y a un rango percentil de 88.

Figura 7–2. Test de dibujo de Goodenough-Harris con calificación de cada reactivo.
Fuente: Reimpreso con autorización de Harris, D. B. (1963). *Children's drawings as measures of intellectual maturity.* Nueva York: Harcourt Brace y World. Copyright © 1963 de The Psychological Corporation. Reproducido con autorización. Derechos reservados.

autoconcepto es un constructo confuso difícil de objetivar. La escala de Sí mismo se ha hecho en gran medida a un lado, aunque algunos psicólogos la emplean simplemente como una extensión no calificable de la entrevista clínica.

La muestra de estandarización del Test de Dibujo Goodenough-Harris fue amplia ($N = 2\,975$ niños), variada en términos geográficos (de áreas urbanas y rurales a lo largo de EUA) y seleccionada de manera cuidadosa para equipararse con los valores de población en EUA en cuanto a ocupación de los padres. La prueba cubre de los 3 a los 16 años de edad, pero las normas son mejores para edades de 5 a 12 años. Más allá de los 12 años, los individuos evaluados comienzan a acercarse a una asíntota de ejecución y se reducen las diferencias según la edad. La escala de Hombre produce una puntuación estándar similar a un CI de desviación, con media de 100 y desviación estándar de 15. La única preocupación es que las normas son ahora bastante anticuadas.

La confiabilidad de la prueba se ha evaluado a través de procedimientos por mitades, estudios *test-retest* y comparaciones intercalificadores (Anastasi, 1975; Frederickson, 1985; Harris, 1963). Las confiabilidades por mitades cercanas a .90 son comunes; no obstante, es poco frecuente que los coeficientes de estabilidad superen al rango de .70, incluso cuando el intervalo *test-retest* es sólo de algunas semanas. Esto sugiere que las calificaciones en el test de Dibujo Goodenough-Harris poseen una considerable banda de error de medición. Por otro lado, la calificación es bastante objetiva. Por lo común, las correlaciones intercalificadores se encuentran en el rango de .90.

Los examinadores que han dominado el elaborado sistema de calificación por puntos pueden utilizar entonces un método global más sencillo denominado Escala de calidad. Esta escala consiste en 24 dibujos (12 para la escala de Hombre y 12 para la de Mujer) utilizados como puntos estandarizados de referencia. El examinador compara los dibujos del individuo examinado con uno de los 12 dibujos de referencia, después consulta un cuadro para determinar la puntuación estándar correspondiente. La calificación de Calidad es más rápida, pero un tanto más rudimentaria: la confiabilidad intercalificadores suele colocarse por debajo del rango de .80.

Con frecuencia, el test Goodenough-Harris se utiliza como medida no verbal de la capacidad cognitiva con niños que tienen trastornos del lenguaje, que pertenecen a minorías o que son bilingües. Oakland y Dowling (1983) consideran que la prueba de dibujo tiene limitada influencia cultural apropiada para la detección inicial con niños de minorías. La prueba funciona mejor con niños pequeños, en particular con aquellos que tienen capacidad intelectual menor (Scott, 1981). Con muestras de niños de cinco años de edad en un centro de cuidado diurno para familias de bajo nivel socioeconómico, Frederickson (1985) informó correlaciones entre el test de Dibujo Goodenough-Harris y el CI Total del WPPSI en el rango de .72 a .80. En otros estudios, las correlaciones con pruebas individuales de CI son más variables, pero la mayor parte están por encima de .50 (Abell, Briesen y Watz, 1996; Anastasi, 1975).

En respuesta a las criticas hacia el test de Dibujo Goodenough-Harris, Naglieri (1988) desarrolló un sistema de calificación cuantitativa y volvió a normar el procedimiento de dibujo de la figura humana. Las normas de su sistema de calificación, el dibujo de la figura humana: un Sistema de Calificación Cuantitativa (DAP), se establecieron con una muestra de 2 622 individuos entre los 5 y 17 años de edad, quienes eran representativos de los datos del censo de EUA para 1980 en cuanto a edad, sexo, raza, región geográfica, grupo étnico, clase social y tamaño de la comunidad. El DAP produce puntuaciones estándar con la conocida media de 100 y desviación estándar de 15. En un estudio con 61 personas entre 6 y 16 años de edad, el DAP se correlacionó en .51 con el CI del WISC-R y produjo calificaciones totales similares con un CI medio de 100 contra una calificación media del DAP de 95 (Wisniewski y Naglieri, 1989). Lassiter y Bardos (1995) encontraron que la calificación de la DAP subestima las calificaciones de CI obtenidas con el WPPSI-R y la K-BIT en una muestra de 50 niños de jardín de niños y primer año.

Quienes han reseñado el DAP lo alaban por su claro sistema de calificación, fuerte confiabilidad y cuidadosa estandarización (Cosden, 1992). No obstante, los resultados de los estudios de validez recomiendan mayor cautela. Harrison y Schock (1994) señalan que la evidencia acumulada con las pruebas del dibujo de la figura humana (DFH) indica una validez predictiva de baja a moderada. A pesar de su popularidad y atractivo, las pruebas de DFH no identifican de manera eficaz a los niños con dificultades de aprendizaje o alteraciones del desarrollo y pueden no ser válidas incluso si se les utiliza como medidas de detección.

PRUEBA HISKEY-NEBRASKA DE APTITUD PARA EL APRENDIZAJE

La *Hiskey-Nebraska Test of Learning Aptitude* (Prueba Hiskey-Nebraska de aptitud para el aprendizaje; H-NTLA) es una escala de ejecución que no requiere del lenguaje para utilizarse con niños entre 3 y 17 años de edad (Hiskey, 1966). Esta prueba puede aplicarse a través de señas y no requiere respuesta verbal de parte de la persona examinada. Sin embargo, pueden emplearse instrucciones verbales con niños normales y con aquellos que tienen alteraciones auditivas leves. La H-NTLA consiste en 12 subpruebas:

Patrones de Cuentas	Patrones con Cubos
Memoria para el Color	Dibujos Incompletos
Identificación de Figuras	Memoria de Dígitos
Asociación de Figuras	Rompecabezas con Cubos
Doblado de Papel	Analogías con Figuras
Nivel de Atención Visual	Razonamiento Espacial

Las puntuaciones naturales de las subpruebas se convierten en un Cociente de Aprendizaje de Desviación (CA) con media de 100 y desviación estándar de 16. Las calificaciones de la H-NTLA se correlacionan mucho con escalas de aprovechamiento desde segundo grado hasta bachillerato (mediana de $r = .49$) y también con el CI de Ejecución del WISC-R ($r = .85$). Aunque el CA rinde calificaciones promedio notablemente cercanas al CI de Ejecución del WISC-R para muestras de niños con alteraciones auditivas y sordos, las calificaciones de la H-NTLA son mucho más variables (Watson y Goldgar, 1985; Phelps y Ensor, 1986). Así, la utilización de la H-NTLA puede aumentar el riesgo de clasificaciones equivocadas por falsos positivos —etiquetar a los niños como dotados cuando nada más son inteligentes o con retraso cuando sólo son limítrofes.

La H-NTLA es adecuada para niños sordos, con alteraciones de habla o lenguaje, retraso mental o bilingües. Un aspecto interesante de esta prueba es el desarrollo de normas paralelas: la H-NTLA se estandarizó con 1 079 niños sordos y 1 074 niños con audición normal entre los 2 años y medio y los 17 años y medio. No obstante, la principal debilidad del instrumento es la deficiencia de estas normas. Por ejemplo, en su mayoría se desconoce la representatividad de la muestra de individuos sordos —elegidos de manera oportunista de una escuela para sordos. La estandarización de la muestra con audición normal se basó en el nivel ocupacional de los padres, según el censo de 1960 para EUA. Sería bastante útil una reestandarización contemporánea y más detallada de la prueba.

PRUEBA DE INTELIGENCIA NO VERBAL-3

La prueba de inteligencia no verbal-3 (TONI-3) es una medida de la capacidad cognitiva que no requiere del lenguaje y que se diseñó para poblaciones discapacitadas o minoritarias (Brown, Sherbenou y Johnsen, 1998). En particular, los autores recomiendan la prueba para la evaluación de personas con afasia, individuos que no hablan inglés, personas con alteraciones auditivas y para aquellos que han experimentado una variedad de traumas neurológicos graves. El examinador proporciona las instrucciones de la prueba con señas y el individuo

responde señalando 1 de 4 o 6 posibles respuestas. La prueba incluye dos formas equivalentes con 50 reactivos de solución de problemas abstractos/figurativos. Éstos se seleccionaron de manera cuidadosa de un conjunto inicial de reactivos según correlaciones reactivo-total, nivel apropiado de dificultad y aceptabilidad para los posibles usuarios y expertos técnicos. Los reactivos de la TONI-3 caen dentro de varias categorías, que incluyen las siguientes:

- equiparación simple
- analogías
- clasificación
- intersección
- progresiones

Excepto por los de equiparación simple, los reactivos de la TONI-3 requieren que la persona examinada resuelva los problemas identificando las relaciones entre figuras abstractas. Muchos de los reactivos son similares en formato a los encontrados en Matrices Progresivas de Raven. La prueba produce dos tipos de puntuaciones: rangos percentiles y cocientes de la TONI-3 (media de 100 y desviación estándar de 15).

La TONI-3 se estandarizó de manera cuidadosa con más de 3 000 personas desde los 6 hasta los 89 años. Las características de la muestra son paralelas a los datos del censo en cuanto a sexo, raza, origen étnico, residencia urbana-suburbana-rural, grado escolar, educación/ocupación de los padres y región geográfica. Los datos de confiabilidad son bastante satisfactorios, con coeficientes de consistencia interna que suelen superar al .90 y confiabilidad de formas alternas en el rango de .80 a .95.

Los estudios de validez de la TONI-3 son pocos, pero la investigación con ediciones anteriores (que tienen un contenido muy similar) apoya a esta prueba como índice de la inteligencia general sin influencia de la cultura. No obstante, los resultados no respaldan la designación de la TONI-3 como una prueba no verbal, excepto en el sentido trivial de que no se requieren respuestas verbales. Por ejemplo, el **Manual** de la TONI-2 presenta coeficientes de correlación en el rango de .70 entre las calificaciones de esta prueba y la subprueba de Artes lingüísticas de la Serie de rendimiento SRA. En general, la investigación con las antecesoras de la TONI-3 indica que es una buena medida de la inteligencia general, pero no sustenta la perspectiva de que es principalmente una medida de inteligencia **no verbal** (Murphy, 1992). En general, la TONI-3 se considera, en gran medida, como un instrumento breve de detección que no emplea el lenguaje y que es adecuada para personas con alteraciones en cuanto a sus capacidades lingüísticas (p. ej., personas que no hablan inglés, individuos con retraso mental, sordos o afásicos). La prueba se estandarizó con más cuidado que la mayor parte y posee excelente confiabilidad. Una característica útil de la TONI-3 es que su aplicación sin límite de tiempo pocas veces excede los 20 minutos.

Dos instrumentos que se analizaron antes en este texto también pueden concebirse como pruebas que no requieren del lenguaje. Matrices Progresivas de Raven y la *Culture Fair Intelligence Test* (Prueba de Inteligencia Culturalmente Justa) de Cattell utilizan reactivos no verbales y no requieren en esencia de interacciones basadas en el lenguaje entre el examinador y la persona evaluada. Una prueba nueva y prometedora que no emplea el lenguaje es la *Universal Nonverbal Intelligence Test* (Prueba Universal de Inteligencia No Verbal; UNIT), una medida amplia y multidimensional de la inteligencia no verbal (McCallum y Bracken, 1997; Reed y McCallum, 1995). Esta prueba se diseñó para niños que tienen deterioro auditivo o limitaciones en sus habilidades en el idioma inglés. La UNIT proporciona una buena medida de *g* y varias subpuntuaciones que incluyen calificaciones claras, basadas en factores, para memoria y razonamiento.

PRUEBAS QUE NO REQUIEREN LA LECTURA Y PRUEBAS CON REDUCCIÓN MOTORA

Como el lector recordará, las pruebas que no requieren lectura se diseñaron en EUA para

individuos iletrados que, sin embargo, pueden comprender el inglés hablado de manera suficientemente correcta como para seguir instrucciones verbales. Las pruebas de inteligencia que no requieren lectura son adecuadas para niños pequeños, personas iletradas e individuos con alteraciones del habla o el lenguaje expresivo; no necesitan ser especializadas o esotéricas; las subpruebas de ejecución de la mayor parte de los instrumentos convencionales pueden considerarse como pruebas que no requieren lectura. Por ejemplo, los examinadores pueden utilizar las subpruebas de ejecución del WISC-III para estimar la inteligencia de individuos con alteraciones del lenguaje.

No obstante, los pacientes con parálisis cerebral u otros padecimientos con alteraciones ortopédicas tendrán calificaciones muy bajas en pruebas que no requieren lectura y necesitan respuestas de manipulación. La obtención de resultados válidos de prueba con dichas personas puede presentar un reto enorme (ejemplo de caso 7–1). Los déficits motores, la mayor tendencia a la fatiga y la falta de exactitud en los movimientos propositivos, comunes a las personas con parálisis cerebral, afectarán de modo negativo su desempeño en las herramientas de evaluación cognitiva. Los clientes con impedimentos ortopédicos necesitan pruebas que no requieran lectura y que tengan reducción motora al mismo tiempo. En particular, las pruebas que permiten una respuesta simple de señalización son las más adecuadas para la evaluación de niños y adultos con parálisis cerebral y otros padecimientos que alteran la motricidad.

PRUEBA PEABODY DE VOCABULARIO CON FIGURAS-III

La *Peabody Picture Vocabulary Test-III* (Prueba Peabody de Vocabulario con Figuras-III, PPVT-III) es la más conocida y de uso más amplio entre las pruebas que no requieren lectura y tienen reducción motora (Dunn y Dunn, 1998). La PPVT-III se utiliza para obtener una medida rápida del vocabulario con personas sordas, pacientes con alteraciones neurológicas e individuos con alteraciones del habla. Aunque la PPVT-III es útil con cualquier persona examinada que no pueda verbalizar bien, la prueba es especialmente útil con individuos que también manifiestan trastornos que alteran la motricidad, como la parálisis cerebral o el evento vascular cerebral.

La PPVT-III tiene dos formas paralelas, cada una de las cuales consiste en 4 láminas de práctica y 204 láminas de prueba. Cada lámina contiene cuatro dibujos lineales de objetos o escenas cotidianas. El examinador presenta una lámina, expresa la palabra verbalmente y le pide a la persona que señale el dibujo que mejor represente la palabra expresada. Los reactivos de prueba se ordenaron según el nivel de dificultad, y se les clasificó en 17 conjuntos de 12 reactivos cada uno para la identificación eficaz de los niveles basal y superior. El nivel de entrada se determina de acuerdo con la edad, y la persona examinada continúa hasta que falla en varios reactivos consecutivos. Aunque la prueba no tiene límite de tiempo, la aplicación pocas veces excede 15 minutos. Las puntuaciones naturales se convierten en equivalentes de edad o puntuaciones estándar (media de 100, desviación estándar de 15).

La PPVT-R se estandarizó con una muestra nacional representativa de varios miles de individuos cuyas edades iban desde dos años y medio hasta 90 años o más. Con base en ediciones anteriores, los coeficientes de confiabilidad por mitades son débiles en el caso de las personas más jóvenes (p. ej., .67 para el nivel de dos años y medio), pero aceptables para individuos mayores (p. ej., .88 para el nivel de 18 años). Para la muestra adulta, la mediana de los coeficientes de confiabilidad por mitades es .82. Las confiabilidades de formas alternas son bastante variables de una muestra a otra (valores de .50 a .90), pero tienden a agruparse alrededor de .80 (Bracken, Prasse y McCallum, 1984; Dunn y Dunn, 1981, Umberger, 1987).

Varias líneas de evidencia apoyan la validez de la PPVT-R, pero sólo como una medida estrecha del vocabulario, no como una medida general de la inteligencia (Alpeter, 1989; Alpeter y Johnson, 1989). Dunn y Dunn (1981) bus-

EJEMPLO DE CASO 7–1
EL RETO DE LA EVALUACIÓN EN LA PARÁLISIS CEREBRAL

Los desafíos inherentes a las consultas especiales se tipifican claramente en el caso de un cliente con accidente vascular cerebral al que un psicólogo consultor sometió a prueba. El joven individuo estaba confinado de manera absoluta a una silla de ruedas movilizada con baterías, excepto cuando un ayudante de planta lo transfería a una cama o silla. Incluso un observador imparcial tendría que estar de acuerdo en que el cliente no parecía muy capaz, sentado en su silla con una postura encorvada, incapaz de controlar su salivación excesiva, con un brazo doblado hacia afuera en un ángulo extraño. No obstante, a pesar de su discapacidad, había logrado un buen grado de independencia personal. Al utilizar un simple mecanismo de control de palanca, podía conducir su silla hasta la tienda de abarrotes, biblioteca y centro comunitario, donde llevaba a cabo transacciones sencillas al señalar a las palabras y frases apropiadas en un cuaderno de espiral forrado de plástico. Debido a su pobre control motor, las interacciones con este cliente tomaban bastante tiempo. Sin embargo, era muy eficiente en las comunicaciones breves. Éste es un intercambio típico, donde en letras mayúsculas se muestran las respuestas del cliente con su cuaderno.

"Según me informaron, usted tiene una nuevo aparato de comunicación que sintetiza la voz, ¿qué le parece?" ME HIZO DOS PREGUNTAS. "Cierto, apuesto que eso le pasa muchas veces. ¿Tiene usted un aparato de comunicación?" SÍ. "¿Qué piensa de él?" NO ES FÁCIL. "Ahora que ya terminamos con las pruebas, ¿le busco a su chofer?" NO. ESPERO. ÉL VA A REGRESAR.

¿Qué tan inteligente es este cliente? ¿Cuál es su nivel de comprensión verbal? ¿Qué tan bien comprende los conceptos abstractos? Por ejemplo, ¿es capaz de comprender los aspectos esenciales del uso de una microcomputadora como ingresar datos, guardar archivos y utilizar comandos directivos? ¿Puede aprender a programar una microcomputadora? Éstas son precisamente las preguntas de canalización que realizó un orientador vocacional de rehabilitación que contemplaba la perspectiva de enormes gastos —miles de dólares— para comprar un sistema de cómputo para este cliente discapacitado.

Con toda seguridad, sería fácil subestimar el potencial de este joven con graves discapacidades motoras y del lenguaje debido a que, en un sentido muy literal —su inteligencia estaba oculta, atrapada en un cuerpo incapacitado. La tarea del examinador consistía en encontrar la mente capaz dentro del cuerpo discapacitado, un reto que sin duda es formidable. Al utilizar el *Test of Nonverbal Intelligence-2* y la *Peabody Picture Vocabulary Test-Revised*, el examinador determinó que este joven cliente poseía cuando menos una inteligencia promedio y que sería muy probable que aprendiera los elementos fundamentales del procesamiento de datos con una microcomputadora.

caron garantizar la validez de contenido al buscar en el *New Collegiate Dictionary* de Webster todas las palabras cuyos significados pudieran representarse a través de un dibujo. Así, los autores tenían en mente un universo especifico de contenido y los reactivos de la PPVT-R parecen constituir un muestreo equitativo de este dominio. Además, los autores utilizaron complejas técnicas de selección de reactivos con base en el modelo de rasgo latente de Rasch-Wright, como una ayuda para darle validez de constructo a la prueba. Este modelo permite que los investigadores construyan una curva de crecimiento para el rasgo latente que mide la PPVT-R (vocabulario auditivo) y que seleccionen los reactivos que más se acercan a la curva. Al utilizar datos de ensayo y calibración, se trazó la curva de manera repetida en una computadora. Si algún reactivo no se adaptaba al modelo de rasgo latente de Rasch-Wright (cur-

va reactivo-característica demasiado plana o inclinada), se le descartaba.

Al utilizar un elaborado modelo de ecuación estructural, Miller y Lee (1993) demostraron que una edición anterior, la PPVT-R, refleja el verdadero nivel de desarrollo en cuanto a vocabulario. Estos investigadores pudieron predecir razonablemente bien la clasificación de las palabras estímulo de la PPVT con base en las características de complejidad de la palabra (cronología de entrada en el idioma inglés, extensión de la palabra, número de significados diversos, frecuencia de ocurrencia). Las variables predictivas proporcionaron una explicación teórica razonable del orden de las palabras en la PPVT-R, es decir, confirmaron la validez de constructo de la prueba.

Los datos sobre la validez concurrente y predictiva de la PPVT-R son un tanto limitados, pero prometedores. Varios investigadores han correlacionado la PPVT-R con medidas de aprovechamiento, donde son comunes las relaciones bajas (r de .30 a .60). Las correlaciones con el aprovechamiento en lectura tienden a ser mayores que con el aprovechamiento en ortografía y aritmética, lo cual sugiere que la PPVT-R tiene adecuada validez discriminativa (Vance, Kitson y Singer, 1985).

Varios investigadores han correlacionado las versiones anteriores de la prueba con medidas de inteligencia, en particular el WISC-R y el WAIS-R, donde la regla han sido las correlaciones altas (cerca de .70) (p. ej., Haddad, 1986; Naglieri y Yazzie, 1983). Como podría esperarse, las correlaciones tienden a ser más altas con el CI Verbal que con el de Ejecución.

En un estudio muy importante e ingenioso, Maxwell y Wise (1984) investigaron la carga de vocabulario de la Peabody en una muestra de 84 pacientes internos en pabellones de psiquiatría y psicología. Su estudio empleó la PPVT, pero esta versión anterior no revisada es tan similar a la PPVT-III que las conclusiones son pertinentes en este caso. Los investigadores estudiaron la hipótesis de que la PPVT evalúa algo más que el vocabulario en el caso de adultos. Además de esta prueba, los investigadores obtuvieron datos del WAIS-R, la Escala Wechsler de memoria, velocidad de escritura del nombre y años de educación. La velocidad de escritura del nombre es simplemente el número de segundos requeridos para que la persona examinada escriba su propio nombre. Aunque todas las variables tuvieron correlaciones significativas con el CI de la PPVT, Vocabulario del WAIS-R tuvo, con mucho, la correlación más alta (r = .88). Lo más importante es que cuando se eliminó la varianza debida a Vocabulario, ninguna de las variables restantes tuvo alguna relación predictiva con la PPVT. En pocas palabras, la Peabody es una buena medida de vocabulario (en particular auditivo), pero puede ser engañosa si se le utiliza como una medida global del intelecto.

La PPVT-III es una revisión muy reciente, de modo que no existe investigación independiente de la prueba. Una advertencia con las ediciones previas es que las puntuaciones estándar de la PPVT-R pueden ser sustancialmente más bajas que los CI de Wechsler, en particular con personas que tienen retraso e individuos que pertenecen a minorías. En una muestra de 21 adultos con retraso mental leve, Prout y Schwartz (1984) encontraron que las puntuaciones estándar de la PPVT-R (media de 56) eran, en promedio, 9 puntos menores que el CI del WAIS-R (media de 65). Naglieri y Yazzie (1983) encontraron una enorme diferencia de 26 puntos con una muestra de niños indígenas navajo que promediaron una puntuación estándar de 61 en la PPVT-R, en contraste con un CI de 87 en el WISC-R.

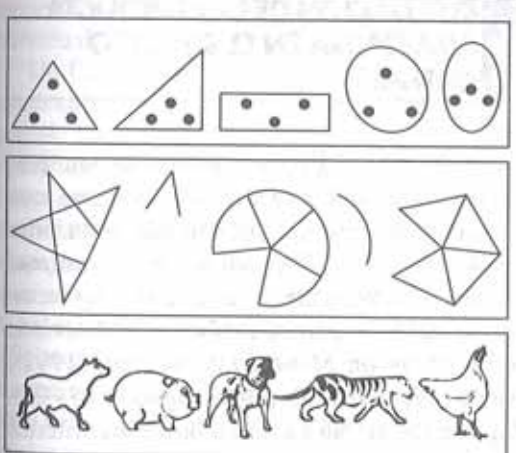
En general, se puede concluir que la PPVT-R es una medida del vocabulario auditivo que posee buenas normas y que es útil con personas que tienen alteraciones motoras o que no saben leer. No obstante, el instrumento no es sustituto de una prueba de la inteligencia general y de ediciones previas y puede subestimar en gran medida el funcionamiento intelectual de personas que pertenecen a minorías o que tienen retraso mental.

ESCALA COLUMBIA DE MADUREZ MENTAL

La *Columbia Mental Maturity Scale* (Escala Columbia de madurez mental; CMMS) es una

prueba sin límite de tiempo de la capacidad de razonamiento, diseñada para niños pequeños con deficiencias sensoriales, motoras o del habla (Burgemeister, Blum y Lorge, 1972). Esta prueba, que ahora se encuentra en su tercera edición, se diseñó originalmente para niños pequeños con parálisis cerebral. La CMMS ha encontrado amplia aceptación con otras discapacidades, incluyendo sordera e impedimentos del habla. El formato de la prueba es claramente simple: de un grupo de 3, 4 o 5 dibujos, el niño tiene que "señalar aquel que no pertenece a los demás". Así, la CMMS no depende de las habilidades de lectura. Las normas de la prueba se establecieron para niños de 3 años y medio a 10 años de edad.

Las 92 láminas de la CMMS examinan discriminación perceptual (p. ej., color, forma, tamaño, uso, número, partes faltantes), clasificaciones y manipulación abstracta de conceptos simbólicos. La prueba requiere de 15 a 20 minutos para su aplicación. Las puntuaciones naturales se convierten en puntuaciones de desviación por edad con media de 100 y desviación estándar de 16. También se puede obtener una puntuación equivalente de edad llamada Índice de Madurez. En la figura 7–3 se presentan varios reactivos característicos de la CMMS.



Nota: Se indica a la persona evaluada que señale al reactivo que no pertenece a los demás.

Figura 7–3. Ejemplos de reactivos característicos de la CMMS.

La muestra de estandarización consiste en 2 600 niños seleccionados de 25 estados, a fin de reproducir los datos del censo de 1960 en cuanto a región geográfica, raza, ocupación de los padres, edad y sexo. La confiabilidad del instrumento es buena, con coeficientes *test-retest* y por mitades que se colocan en la parte superior del rango de .80. La validez concurrente del instrumento es adecuada: las correlaciones con otros instrumentos se encuentran por lo común en el rango del .50. El Manual presenta una correlación de .67 con la Stanford-Binet: Forma L-M.

La CMMS tiene dos graves deficiencias. La primera es el hecho obvio de que las normas, que se obtuvieron en el decenio de 1960, son ahora extremadamente obsoletas. Con base en lo que se ha observado con otras pruebas anticuadas, es probable que la CMMS sobreestime el funcionamiento de los niños actuales (véase el análisis sobre el efecto Flynn en el siguiente tema). Otro problema es que cuando la prueba se aplica a niños muy pequeños, la escala es muy vulnerable a las conjeturas aleatorias. Kaufman (1978) señala que un niño pequeño que simplemente adivina obtendrá, en promedio, una puntuación de 82. Con un poco de suerte en sus conjeturas, un niño de cinco años de edad con retraso mental podría obtener una calificación dentro del promedio. Así, la CMMS tiene un valor limitado en la evaluación de niños muy pequeños de los que se sospecha discapacidad intelectual. No obstante, las conjeturas tienen un efecto mínimo en el caso de niños mayores, debido al mayor número de alternativas y a la mayor cantidad de reactivos que suelen aplicarse. A pesar de estas limitaciones, la CMMS es una medida útil de la capacidad de razonamiento, que no requiere del lenguaje y que es adecuada para niños en edad escolar, pero los nuevos datos normativos son muy necesarios.

PRUEBAS PARA INDIVIDUOS CON ALTERACIONES VISUALES

En la evaluación del funcionamiento intelectual de personas con alteraciones visuales, los exa-

minadores han dependido históricamente de adaptaciones de la Stanford-Binet. La revi- sión Hayes-Binet para el examen de individuos con alteraciones visuales se basó en la Stanford-Binet de 1916; desde entonces, este instrumento ha atravesado por varias revisiones. La adaptación más reciente es la Perkins-Binet (Davis, 1980), que conserva la mayor parte de los reactivos verbales de la Stanford-Binet, pero también adapta otros a un modo táctil. La Perkins-Binet posee una confiabilidad aceptable por mitades y muestra altas correlaciones con las escalas verbales del WISC-R (Coveny, 1972; Teare y Thompson, 1982). Los creadores de la Perkins-Binet han reconocido que los problemas visuales existen en un continuo y, por tanto, han desarrollado normas para niños con visión utilizable (Forma U) y no utilizable (Forma N).

Los creadores de pruebas también han tenido éxito en modificar las Escalas de Ejecución de Wechsler para utilizarlas con individuos que tienen alteraciones visuales. La Escala háptica de inteligencia para adultos invidentes (del inglés *Haptic Intelligence Scale for Adult Blind*; HISAB) consiste en seis subpruebas, cuatro de las cuales se asemejan a las de Símbolos y Dígitos, Diseño con Cubos, Composición de Objetos y Figuras Incompletas de la Escala de Ejecución del WAIS (Shurranger, 1961; Shurrager y Shurrager, 1964). Las otras dos subpruebas son Aritmética con Cuentas, que implica el uso de un ábaco para resolver problemas aritméticos, y Tablero de Patrones, que requiere que el examinado reproduzca el patrón que ha sentido en un tablero que tiene filas de orificios con clavijas en ellos. La confiabilidad de la HISAB es excelente y los autores proporcionan datos normativos sobre una muestra de adultos con discapacidad visual. Lo más alentador es que las calificaciones de la HISAB se correlacionan en .65 con el CI Verbal del WAIS (Shurrager y Shurrager, 1964).

Otro instrumento interesante es la *Blind Learning Aptitude Test* (Prueba de aptitud para el aprendizaje para ciegos; BLAT), una prueba táctil para niños invidentes de los 6 a 16 años de edad (Newland, 1971). Los reactivos de la BLAT se encuentran en forma de bajorrelieve que consiste en puntos y rayas similares al sistema Braille. Los reactivos son de seis tipos diferentes: reconocimiento de diferencias, reconocimiento de semejanzas, identificación de progresiones, identificación del elemento faltante en una matriz de 2 × 2, terminado de una figura incompleta e identificación del elemento faltante en una matriz de 3 × 3. La mayor parte de los reactivos se adaptaron de Matrices Progresivas de Raven y de la *Culture Fair Intelligence Test* de Cattell. La BLAT se estandarizó con 760 niños con alteraciones visuales, pero las normas son anticuadas y el manual de la prueba es incompleto y un tanto descuidado (Herman, 1988). No obstante, la prueba posee una confiabilidad excepcional y se correlaciona muy bien con la Hayes-Binet ($r = .74$) y con la Escala Verbal del WISC ($r = .71$). En conjunto con una prueba verbal, la BLAT es un instrumento prometedor para probar la inteligencia de niños con discapacidades visuales. Sin embargo, la prueba se beneficiaría sustancialmente de revisiones menores, actualización de las normas y de un manual de prueba más completo.

EVALUACIÓN DE LA CONDUCTA ADAPTATIVA EN EL RETRASO MENTAL

La evaluación del retraso mental es una preocupación compleja y multifacética que, con todo derecho, merece un capítulo o un libro aparte. Debido a las limitaciones de espacio, aquí se tratará brevemente, y se referirá a los lectores interesados a los trabajos de la *American Association on Mental Retardation* (1992), Nihira (1985) y Sattler (1988, capítulos 15 y 21). Aquí se resumirán en forma breve los criterios diagnósticos para el retraso mental y después se revisarán, específicamente, dos instrumentos contrastantes de evaluación. El texto termina con un resumen tabular de varias medidas notables de la conducta adaptativa.

DEFINICIÓN DE RETRASO MENTAL

La fuente con mayor autoridad para la definición de retraso mental es el manual de terminología y clasificación de la *American Association on Mental Retardation* (AAMR, 1992). Este manual define al **retraso mental** como sigue:

> Retraso mental se refiere a las limitaciones sustanciales en el funcionamiento actual. Se caracteriza por un funcionamiento intelectual que se encuentra de manera significativa por debajo del promedio y que existe de modo concurrente con limitaciones relacionadas en dos o más de las siguientes áreas de habilidades adaptativas aplicables: comunicación, autocuidado, vida en el hogar, habilidades sociales, uso de la comunidad, independencia, salud y seguridad, conocimientos académicos funcionales, descanso y trabajo. El retraso mental se manifiesta antes de los 18 años de edad (AAMR, 1992).

El Manual también especifica que el funcionamiento intelectual que se encuentra de manera significativa por debajo del promedio es un CI de 70 a 75, o menos, en escalas con una media de 100 y desviación estándar de 15. En pruebas como la Stanford-Binet: Cuarta Edición, que posee una desviación estándar de 16, el rango promedio de funcionamiento intelectual deficiente sería un CI de 68 a 73 o menor. El Manual también afirma de modo explícito la importancia del juicio profesional en los casos individuales.

Un CI bajo no es, en sí mismo, un fundamento suficiente para el diagnóstico de retraso mental. La definición de la AAMR también especifica un segundo criterio, el de las limitaciones en dos o más áreas pertinentes de habilidad adaptativa. Un diagnóstico de retraso mental se justifica sólo cuando un individuo presenta un CI suficientemente bajo y limitaciones en habilidades de adaptación. Además, estas deficiencias en intelecto y en funcionamiento adaptativo deben haber surgido durante el periodo del desarrollo —definido como aquél entre el nacimiento y los 18 años de edad.

Este **Manual** más reciente de la AAMR representa un distanciamiento con respecto a la terminología anterior que reconocía cuatro niveles de retraso: leve, moderado, grave y profundo. En vez de enfocarse en las deficiencias de la persona, el **Manual** introduce una jerarquía de "Intensidades de los apoyos requeridos", que redirige la atención hacia las necesidades de rehabilitación del cliente. Los cuatro niveles de apoyo requerido son: intermitente, limitado, extenso y generalizado. No obstante, es probable que la terminología anterior referente a los niveles de retraso se conserve por bastante tiempo, de modo que en el cuadro 7–1 se ha elegido una mezcla del viejo enfoque con el nuevo. El lector observará que existe una zona de incertidumbre entre los niveles de retraso, lo que significa que se requiere del juicio clínico acerca de todas las fuentes de información para elaborar el diagnóstico. Lo que es más,

Cuadro 7–1. Cuatro niveles de retraso mental

Retraso mental leve: CI de 50 a 55 y de 70 a 75+, se requiere **apoyo intermitente**. Habilidades sociales y de comunicación razonables; con educación especial, puede lograr un nivel de sexto grado para finales de la adolescencia; alcanza un nivel social y vocacional adecuados con entrenamiento especial y supervisión; independencia parcial en disposiciones de vivienda.

Retraso mental moderado: CI de 35 a 44 a 50 a 55, se requiere **apoyo limitado**. Habilidades sociales y de comunicación adecuadas, pero poca conciencia de sí mismo; con educación especial ampliada logra un nivel de cuarto grado; funciona en un taller protegido, pero necesita supervisión en las disposiciones de vivienda.

Retraso mental grave: CI de 20 a 25 a 35 a 40, se requiere **apoyo profuso**. Pocas o ninguna habilidad de comunicación; alteraciones sensoriales y motoras; no se beneficia del entrenamiento académico; entrenable en hábitos básicos de salud.

Retraso mental profundo: CI inferior a 20 a 25, se requiere **apoyo generalizado**. Funcionamiento mínimo; incapaz de cuidar de sí mismo; necesita constante atención y supervisión profesionales.

Fuente: Basado en AAMR (1992) y Patton, Payne y Beirne-Smith (1986).

aunque estos niveles se calibran por medio de los rangos de CI, se recordará al lector que la persona evaluada también debe mostrar un déficit correspondiente en dos o más áreas de habilidad adaptativa. Bajo ninguna circunstancia una prueba de CI es base suficiente para el diagnóstico de retraso mental.

Las limitaciones en habilidades adaptativas son más difíciles de confirmar que un CI bajo. El **Manual** de la AAMR lista 10 áreas diferentes de habilidad adaptativa y especifica que el cliente debe mostrar limitaciones sustanciales en dos o más de ellas:

- comunicación
- autocuidado
- vida en el hogar
- habilidades sociales
- uso de la comunidad
- independencia
- salud y seguridad
- conocimientos académicos funcionales
- descanso
- trabajo

En cuanto a la manera en que han de evaluarse estas limitaciones, el **Manual** propone que son deseables las medidas bien normadas de las habilidades adaptativas, pero la determinación final siempre es una cuestión de juicio profesional.

Un autor de pruebas enfrenta importantes problemas para medir las limitaciones en conducta adaptativa. Casi el único hecho sólido que se tiene en este campo es que las expectativas ambientales hacia la conducta adaptativa aumentan de manera pronunciada desde el nacimiento hasta la adultez joven. Además, la expresión de la conducta adaptativa cambia de carácter a lo largo de este periodo. En la infancia, las conductas adaptativas pueden reflejarse en las habilidades sensoriomotoras y en la facilidad para el lenguaje. En la adultez, los logros vocacionales y la responsabilidad social se vuelven importantes. Igual que ocurre con la evaluación intelectual, las herramientas para valorar la conducta adaptativa deben estar cuidadosamente graduadas de acuerdo con la edad.

El primer instrumento estandarizado para evaluar la conducta adaptativa fue la *Vineland Social Maturity Scale* (Escala Vineland de Madurez Social; Doll, 1935, 1936). Un tanto simplista y cruda para las normas modernas, la escala Vineland original consistía en 117 reactivos separados ordenados en un formato de escala por años de edad. Un informante familiarizado con la persona examinada marcaba los reactivos aplicables. A partir de estos resultados, el examinador calculaba un equivalente de edad social, útil en el diagnóstico de retraso mental. Esta escala todavía es un instrumento respetado que ha atravesado por varias revisiones y ahora se conoce como *Vineland Adaptative Behavior Scales* (Escalas Vineland de Conducta Adaptativa; Sparrow, Balla y Ciccetti, 1984).

Desde la aparición de la Escala Vineland original se han publicado más de 100 escalas de conducta adaptativa (Nihira, 1985; Reschly, 1990; Walls, Werner, Bacon y Zane, 1977). Estos instrumentos varían en gran medida en cuanto a estructura, propósito y población blanco. En términos generales, se pueden distinguir dos tipos de instrumentos diseñados para dos propósitos diferentes. Un grupo lo integran principalmente las escalas referidas a normas que se utilizan en general para asistir en el diagnóstico y clasificación. Otro grupo está formado por escalas referidas a criterio y se utiliza de manera principal para asistir en el entrenamiento y la rehabilitación. Se ha elegido un instrumento representativo de cada grupo para realizar un análisis más detallado.

ESCALAS DE CONDUCTA INDEPENDIENTE-REVISADAS

Las Escalas de Conducta Independiente-Revisadas (SIB-R; del inglés *Scales of Independent Behavior-Revised*; Bruininks, Woodcock, Weatherman y Hill, 1996) son una ambiciosa medida multidimensional de la conducta adaptativa que es de suma utilidad para la evaluación del retraso mental. El instrumento consiste en 259 reactivos de conducta adaptativa que se organizaron en 14 subescalas. La escala se com-

pleta con la ayuda de un padre, maestro o encargado del cuidado del paciente que esté bien familiarizado con las conductas cotidianas de la persona examinada. Para cada subescala, el examinador lee una serie de reactivos y para cada uno registra una calificación de 0 (nunca o casi nunca realiza la tarea) a 3 (realiza muy bien la tarea). Una característica útil de la SIB-R es que el examinador necesita un entrenamiento y experiencia mínimos. Por supuesto, se requiere un nivel mucho más alto de competencia para evaluar los resultados y tomar decisiones sobre asignación o tratamiento.

Las 14 subescalas de la SIB están ordenadas en cuatro agrupamientos, como se describe en el cuadro 7–2. A su vez, estos cuatro agrupamientos constituyen la Escala de Independencia General. Cada subescala incluye un pequeño número de reactivos separados, ordenados con base en el desarrollo. Por ejemplo, la subescala de Alimentación y Preparación de Comidas tiene 19 reactivos por grados, que incluyen: ensarta la carne con un tenedor, come la sopa con cuchara, toma porciones de tamaño apropiado y prepara bocadillos que no requieren cocinarse. Para cada subescala, los reactivos se aplican hasta llegar a un determinado nivel superior (p. ej., 3 a 5 reactivos consecutivos calificados con 0).

Las puntuaciones naturales de cada subprueba se suman para obtener una calificación parcial. Después se suman las calificaciones parciales de cada agrupamiento para obtener una calificación por agrupamiento. La puntuación de la Escala de Independencia General se deriva del promedio de las cuatro calificaciones de agrupamiento. Las calificaciones de subprueba, agrupamiento e Independencia General pueden convertirse después en una variedad de calificaciones normativas que permiten la comparación del desempeño del examinado con el de una muestra normativa nacional. Las escalas normativas incluyen: calificaciones por edad, rangos percentiles, puntuaciones estándar, estaninas y equivalentes de curva normal.

Una parte autónoma y única de la SIB-R también evalúa la conducta desadaptativa al medir la frecuencia y gravedad de las conductas problema. La Escala de Conductas Problema incluye ocho categorías principales de desadaptación personal y social que pueden afectar la conducta adaptativa: Daños contra sí mismo, Daños contra los demás, Destrucción de la propiedad, Conducta indisciplinada, Hábitos inusuales y repetitivos, Conducta socialmente ofensiva, Conducta de retraimiento o falta de atención y Conducta no cooperativa. Se proporcionan listas de conductas problema y la persona que responde debe indicar las conductas exhibidas por el individuo evaluado. Además, quien responde a la escala describe la conducta más grave en cada categoría y la califica según la frecuencia de ocurrencia, gravedad y manejo típico.

La estandarización de la SIB-R se concibió y ejecutó de manera adecuada. La muestra normativa incluyó a 2 182 personas en una muestra que refleja las características del censo de 1990. Los datos normativos abarcan personas desde los 3 meses hasta adultos de más de 80 años de edad. También se sometió a prueba a una muestra adicional de más de 1 000 individuos con retraso mental, discapacidades para el aprendizaje, trastornos de conducta y alteraciones auditivas. El valor de la SIB-R se fortaleció aún más al vincular dichas escalas con las normas de la Batería Psicoeducativa Woodcock-Johnson-Revisada. La SIB-R es un componente de esta batería más amplia, pero se le puede utilizar de manera independiente.

La confiabilidad de la SIB-R se considera respetable en términos generales, pero es variable de una subescala a otra y de un grupo de edad a otro. Las subescalas individuales tienden a mostrar confiabilidades por mitades cercanas a 0.80; los cuatro agrupamientos tienen confiabilidades mixtas medianas alrededor de .90; la Escala de Independencia General tiene una confiabilidad muy fuerte de .90 (Bruininks, Woodcock, Weatherman y Hill, 1996).

Los datos iniciales de validez para la SIB-R son muy prometedores. Por ejemplo, las calificaciones medias para diversas muestras de personas con discapacidad y sin ella muestran relaciones confirmatorias: las puntuaciones de la SIB-R son más bajas entre aquellos individuos de los que se sabe que tienen la discapacidad

Cuadro 7–2. Subescalas y Agrupamientos de las Escalas de Conducta Independente-Revisada

1. Habilidades motoras

Motoras gruesas: 19 habilidades de músculos grandes, como sentarse sin apoyo o tomar parte en actividades físicas vigorosas.

Motoras finas: 19 habilidades de músculos pequeños, como levantar objetos pequeños o armarlos.

2. Habilidades sociales y de comunicación

Interacción social: 18 habilidades que requieren interacción con otras personas, como pasar juguetes a otros individuos o hacer planes con amigos para asistir a actividades sociales.

Comprensión del lenguaje: 18 habilidades que implican la comprensión de lenguaje hablado o escrito, como dirigir la mirada hacia el hablante o leer.

Expresión del lenguaje: 20 habilidades que implican hablar, como hacer sonidos para obtener atención o explicar un contrato por escrito.

3. Habilidades de la vida personal

Alimentación y preparación de comida: 19 habilidades relacionadas con la alimentación y preparación de comidas, que van desde beber en un vaso hasta planear una comida.

Control de esfínteres: 17 habilidades necesarias para uso del baño y retrete.

Vestido: 18 habilidades relacionadas con vestirse, que van desde levantar los brazos y piernas mientras se le viste hasta hacer arreglos para alterar la ropa.

Autocuidado personal: 16 tareas que se relacionan con el arreglo personal básico y la conservación de la salud; por ejemplo, lavarse las manos y hacer una cita con el médico.

Habilidades domésticas: 18 tareas necesarias para mantener en orden una casa, que van desde colocar los platos vacíos en el fregadero, hasta seleccionar una vivienda apropiada.

4. Habilidades de la vida en comunidad

Tiempo y puntualidad: 19 tareas que implican conceptos de tiempo y manejo del mismo, como cumplir con una cita.

Dinero y valores: 20 habilidades relacionadas con conceptos de dinero, como ahorro de dinero y uso de crédito.

Habilidades laborales: 20 habilidades relacionadas con hábitos prevocacionales y laborales; por ejemplo, indicar que se ha terminado con una tarea asignada.

Orientación en hogar y comunidad: 18 habilidades implicadas en desplazarse dentro del hogar y vecindario y trasladarse dentro de la comunidad; por ejemplo, localizar a un dentista.

más grave en aprendizaje y adaptación. Para los individuos discapacitados, las puntuaciones de la SIB-R se correlacionan en gran medida con calificaciones de inteligencia (.80), mientras que en el caso de individuos examinados que no tienen discapacidad la relación es mínima (Bruininks, Woodcock, Weatherman y Hill, 1996).

En resumen, la SIB-R es una excelente herramienta para proporcionar discernimientos acerca del nivel actual de funcionamiento de la persona en situaciones reales en el hogar, escuela y entornos comunitarios. Aunque este instrumento no tiene una correspondencia uno a uno con las 10 áreas de habilidad adaptativa listadas en la definición de retraso mental, existe una semejanza sustancial. Por ejemplo, las siguientes áreas de las habilidades adaptativas listadas por la AAMR se cubren de manera adecuada en las subescalas o agrupamientos de la SIB-R: comunicación, autocuidado, vida en el hogar, habilidades sociales, uso de la comunidad, salud y seguridad, y trabajo. La SIB-R o un instrumento similar se clasifica como complemento obligatorio de las pruebas individuales de inteligencia en el diagnóstico y evaluación del retraso mental.

LISTA DE VERIFICACIÓN DE LA CONDUCTA DE VIDA INDEPENDIENTE (ILBC)

La *Independent Living Behavior Checklist* (Lista de Verificación de la Conducta de Vida Independiente; ILBC) es una extensa lista de 343 habilidades de vida independiente, clasificadas y presentadas en seis categorías: movilidad, autocuidado, mantenimiento y seguridad del hogar, alimentación, social y comunicación y académica funcional (Walls, Zane y Thvedt, 1979). A diferencia de la mayor parte de los instrumentos analizados hasta ahora en este texto, la ILBC carece por completo de relación con normas. El único propósito de esta prueba es facilitar el entrenamiento del individuo examinado en las habilidades requeridas para la vida independiente. Para este propósito, un conjunto de habilidades cuidadosamente seleccionadas referidas a criterio funciona mejor que un grupo de calificaciones con base en normas. La ILBC se enfoca en aquello que la persona puede hacer, no en cómo se compara dicha persona con otras. No se especifica un rango exacto de edad, pero el instrumento parece adecuado para individuos desde los 16 años hasta la edad adulta.

Para cada habilidad, la ILBC especifica una condición, una conducta y un estándar. El cuadro 7–3 lista una muestra de reactivos de la ILBC. El lector observará que los tres componentes (condición, conducta y estándar) se definen con la suficiente precisión como para que sea posible que observadores razonables coincidan acerca de cuándo se ha dominado una habilidad. De hecho, la confiabilidad *test-retest* y la concordancia entre observadores para las habilidades de la ILBC van de .96 a un perfecto 1.00.

Los reactivos dentro de cada categoría de la ILBC se seleccionaron con cuidado para englobar las habilidades importantes y pertinentes para la vida independiente. En apariencia, los autores tuvieron éxito en identificar las habilidades esenciales, dado que su instrumento tiene una superposición de 100% con otra lista de verificación de la vida independiente, inicialmente

Cuadro 7–3. Muestra de reactivos de la ILBC

Espátula de plástico 35
Condición: Dar un tazón que contiene ingredientes, una cacerola y una espátula de plástico.
Conducta: El cliente vierte los ingredientes en la cacerola y raspa los lados del tazón con la espátula.
Estándar: Conducta dentro de un periodo de dos minutos. No deben derramarse los ingredientes. Todos deben retirarse del tazón.
Elogios 30
Condición: Dar un *role play* o situación natural en la que se elogia al cliente.
Conducta: El cliente acepta el(los) elogio(s) (p. ej., dice "Gracias").
Estándar: En el *role play* o situación natural, todas las personas entrevistadas deben expresar de manera independiente que el cliente aceptó el elogio de manera amable y no fue demasiado cortés o vanidoso.
Direcciones 38
Condición: Dar un trozo de papel con la dirección de un lugar localizado dentro de una distancia de tres cuadras con respecto al cliente.
Conducta: El cliente encuentra el lugar apropiado con o sin ayuda.
Estándar: Conducta en un periodo de una hora. Se debe encontrar el lugar apropiado. El cliente puede encontrar el lugar por sí solo o con ayuda (p. ej., pidiendo indicaciones a otras personas, como un policía).

Fuente: Reproducido con autorización de Walls, R. T., Zane, T. y Thvedt, J. E. (1979). *The Independent Living Behavior Checklist.* Dunbar, WV: West Virginia Research and Training Center.

desconocida (Schwab, 1979). Además, los reactivos de la ILBC se ordenaron cuidadosamente del más fácil al más difícil. Cuando se le utiliza de manera continua a lo largo de un periodo de entrenamiento de varios años, la ILBC proporciona, al mismo tiempo, una lista de verificación para las habilidades dominadas y una guía para mayor rehabilitación.

MEDIDAS ADICIONALES DE LA CONDUCTA ADAPTATIVA

Se recuerda al lector que las medidas de conducta adaptativa varían en gran medida. Algunas escalas están diseñadas principalmente para el diagnóstico y otras para la rehabilitación; algunas son útiles con personas que tienen retraso mental grave y profundo y que nunca obtendrán un empleo; otras son adecuadas para individuos con retraso mental leve que buscan entrenamiento vocacional. Algunas escalas son útiles exclusivamente con niños, otras, con adultos. Estos instrumentos no son intercambiables y el usuario potencial debe estudiar con cuidado sus fortalezas y limitaciones. En el cuadro 7–4 se ha resumido 1 o 2 características sobresalientes de varias de las medidas más conocidas de conducta adaptativa. Reschly (1990) revisa otras escalas adicionales.

Cuadro 7–4. Medidas de conducta adaptativa

Adaptive Behavior Inventory for Children
(Inventario de Conducta Adaptativa para Niños; ABIC)

Edades 5-0 a 11-11 años

La ABIC es parte del sistema de evaluación multicultural pluralista (SOMPA). Seis áreas de conducta adaptativa miden conducta de rol en el hogar, vecindario y comunidad. Sattler (1988) advierte que la ABIC no hace ninguna disposición en cuanto a considerar las oportunidades del niño para aprender las conductas adaptativas.

Mercer y Lewis (1978)

AAMR Adaptive Behavior Scales —Residential and Community, 2nd Edition
(Escalas de Conducta Adaptativa AAMR —Residencial y Comunitaria, 2a Edición; ABS-RC:2)

Edades de 18-0 hasta 80-0

Este instrumento es una proeza psicométrica que raya en la exageración. La muestra normativa incluye a más de 4 000 personas con discapacidades del desarrollo, que provienen de 43 estados y residen en entornos comunitarios o residenciales. Las confiabilidades de consistencia interna y de *test-retest* superan al .80 en todas las calificaciones de dominio. El instrumento se ha validado de manera extensa y distingue con claridad a las personas clasificadas de manera independiente en diferentes niveles de conducta adaptativa.

Nihira, Leland y Lambert (1993)

AAMR Adaptive Behavior Scales —School, 2nd Edition
(Escalas de Conducta Adaptativa AAMR —Versión escolar, 2a Edición; AAMR —ABS)

Edades de 3-0 a 18-11 años

Este instrumento es útil con preescolares y niños en edad escolar de los que existe la sospecha de retraso mental. Uno de los padres o el encargado del cuidado del niño es quien llena la escala. La escala escolar AAMR actual se divide en dos partes. La primera consta de nueve dominios conductuales que tienen que ver con independencia personal. La segunda parte se enfoca en la valoración de conductas socialmente desadaptativas. La escala produce cinco puntuaciones factoriales al igual que una puntuación total de independencia.

Lambert, Nihira y Leland (1993)

Cuadro 7–4. Medidas de conducta adaptativa (continuación)

Vineland Adaptive Behavior Scales
(Escalas Vineland de conducta adaptativa; VABS)

Edades desde el nacimiento hasta los 19-11 años.

La VABS es un instrumento sumamente respetado que tiene buena validez concurrente; así como correlaciones de .50 a .80 con el WISC-R y la Stanford-Binet. No obstante, algunos reactivos requieren conocimiento que es posible que los informantes no conozcan (p. ej., si el niño dice 100 palabras reconocibles). Silverstein (1986) critica los datos normativos, señalando saltos discontinuos en puntuaciones estándar de un grupo de edad a otro.

Sparrow, Balla y Cicchetti (1984)

TMR School Competency Scales
(Escalas de aptitud escolar para DME; retraso mental entrenable)

Dos formas: edades 5 a 10-11 años y 11 años en adelante.

Diseñada para evaluar las habilidades sociales y personales de niños con retraso mental entrenable en un ambiente escolar. Las cinco escalas incluyen Perceptomotora, Iniciativa-responsabilidad, Cognición, Personal-social y Lenguaje. El maestro elige entre cuatro niveles de aptitud para cada reactivo. En la actualidad faltan datos de validez.

Levine, Elzey, Thormahlen y Cain (1976)

Balthazar Scales of Adaptive Behavior
(Escalas Balthazar de conducta adaptativa)

Edades 5 años hasta la adultez

Un instrumento útil para evaluar la eficacia de programas de entrenamiento con individuos que tienen retraso mental grave y profundo. Las Escalas Balthazar proporcionan un análisis detallado de las conductas de alimentación, vestido y entrenamiento de esfínteres.

Camelot Behavioral Checklist
(Lista de verificación conductual Camelot)

Adultos

La Camelot consta de 399 descripciones conductuales agrupadas en 40 subdominios y 10 dominios. Ejemplos de reactivos incluyen "vierte líquidos", "encera pisos", "puede hervir comida" y "puede realizar trabajos de engrapado". Las 10 calificaciones de dominio producen percentiles referidos a la norma basados en una muestra de 624 personas con discapacidad del desarrollo. La Camelot tiene excelente confiabilidad y se correlaciona de .70 a .86 con las pruebas Wechsler y la Stanford-Binet.

Bruininks-Oseretsky Test of Motor Proficiency
(Prueba de destreza motora Bruininks-Oseretsky)

Edades 4-6 a 14-6 años

Esta prueba especializada se desarrolló para ayudar a educadores y clínicos a identificar y evaluar disfunciones motoras en personas que tienen retraso mental u otras dificultades del desarrollo. Ocho subpruebas examinan la velocidad y agilidad para correr, equilibrio, coordinación bilateral, fortaleza, coordinación de miembros superiores, velocidad de respuesta, control visomotor, y velocidad y destreza en miembros superiores.

Bruininks (1978)

RESUMEN

1. En el decenio de 1970, la renovada preocupación de la sociedad acerca de las necesidades de las personas con discapacidades se tradujo en legislación federal. La Ley Pública 93-112 prohibió la discriminación basada en la discapacidad. La Ley Pública 94-142 decretó que los niños discapacitados en edad escolar recibieran una evaluación apropiada y oportunidades educativas.
2. La Ley Pública 94-142, Decreto de educación para todos los niños con discapacidad, ha tenido impactos sobresalientes en las prácticas de evaluación de los individuos con discapacidad. La ley especifica: evaluación no discriminatoria, examen efectuado en la lengua materna, empleo de pruebas válidas aplicadas por personal con el entrenamiento apropiado y uso de procedimientos auxiliares; por ejemplo, evaluación de salud, audición, visión y estado emocional.
3. La Ley Pública 99-457 requiere que los estados proporcionen educación pública apropiada y gratuita a niños entre los 3 y 5 años de edad que sufran discapacidades. Además, la ley decreta subsidios económicos para los estados que atiendan las necesidades de los lactantes e infantes discapacitados y de sus familias.
4. El Decreto sobre los estadounidenses con discapacidades (ADA) de 1990 ordena que las instituciones deben hacer arreglos razonables a la prueba para las personas con discapacidad, incluyendo disposiciones como aumento en el tiempo para pruebas con límites de tiempo en el caso de personas con discapacidad para el aprendizaje o trastornos relacionados.
5. La *Leiter International Performance Scale-Revised* es una medida sin límite de tiempo de la organización perceptual y de la capacidad de razonamiento. La prueba puede aplicarse completamente a través de señas —el examinando compara pequeñas tarjetas laminadas debajo de las ilustraciones correspondientes en un folleto colocado en forma de atril.
6. El Test de Dibujo Goodenough-Harris es una prueba breve de detección de la inteligencia en la que se anima a la persona evaluada a hacer un buen dibujo de un hombre. Los 73 elementos a calificar incluyen partes de cuerpo, detalles, perspectiva, proporción e implicación de libertad de movimiento. El Test del Dibujo de la Figura Humana de Naglieri (1988) es una versión actualizada de la Prueba de Dibujo.
7. La *Hiskey-Nebraska Test of Learning Aptitude* es una escala de ejecución que no requiere del lenguaje y que se utiliza con niños entre los 3 y 17 años de edad. La prueba se emplea con niños sordos, con alteraciones del habla o lenguaje, retraso mental o bilingües. Sus normas se establecieron originalmente en 1960 y, por tanto, la prueba requiere reestandarización.
8. El Test de Inteligencia no Verbal-3 (TONI-3) es una medida no verbal, de opción múltiple, de la capacidad cognitiva, que se diseñó para poblaciones especiales y se estandarizó cuidadosamente para edades desde los 5 hasta 85 años. La mayor parte de los reactivos requiere que el individuo evaluado identifique las relaciones entre figuras abstractas. La TONI-3 es un buen índice de la inteligencia general —al contrario de la no verbal.
9. La *Peabody Picture Vocabulary Test-III* (PPVT-III) es adecuada para obtener una medida rápida del vocabulario auditivo con personas sordas o discapacitadas (p. ej., por evento vascular cerebral o parálisis cerebral). El examinador dice una palabra y el individuo intenta seleccionar entre cuatro láminas aquella que representa la palabra.
10. La *Columbia Mental Maturity Scale* (CMMS), cuyas normas se establecieron en

el decenio de 1960 para personas de 3 años y medio a 10 años de edad, es una prueba de la capacidad de razonamiento sólida en términos psicométricos y que se diseñó para niños con alteraciones sensoriales, motoras o del habla. Se pide al niño que, de un grupo de 3 a 5 dibujos, "señale el que no pertenece a los demás". La CMMS requiere una nueva estandarización.

11. Pruebas respetadas para personas con alteraciones visuales incluyen la Perkins-Binet, una adaptación de la Stanford-Binet; la HISAB, una modificación de las subpruebas de ejecución de Wechsler; y la *Blind Learning Aptitude Test* (BLAT), una medida de tipo Braille acerca de la formación de conceptos y el razonamiento abstracto.
12. El retraso mental se define con base en tres criterios: funcionamiento intelectual general inferior al promedio, el cual se define por lo común como un CI menor a 70 (o 75 en casos excepcionales), limitaciones en dos o más áreas de habilidad adaptativa y brote previo a los 18 años de edad.
13. Las *Scales of Independent Behavior-R* (SIB-R) son una medida de la conducta adaptativa sumamente útiles en la evaluación del retraso mental. Un padre, maestro o persona encargada del cuidado del paciente responde a una serie de 14 subescalas relacionadas con habilidades motoras, habilidades sociales y de comunicación, habilidades de vida personal y habilidades de vida comunitaria.
14. La *Independent Living Behavior Checklist* (ILBC) es un compendio referido a criterio y originado en el tratamiento que presenta 343 habilidades para la vida independiente. La confiabilidad *test-retest* y la concordancia interobservadores para las habilidades de la ILBC son excepcionalmente altas. Esta prueba ayuda a guiar la rehabilitación continua de personas con discapacidades del desarrollo.

TÉRMINOS Y CONCEPTOS CLAVE

Ley Pública 93-112 p. 264
Ley Pública 94-142 p. 265
Ley Pública 99-457 p. 266
Decreto sobre los estadounidenses con discapacidades p. 266
Retraso mental p. 279

TEMA 7B
Sesgo de una prueba y otras controversias

La cuestión del sesgo de una prueba
Valores sociales y equidad de una prueba
Determinantes genéticos y ambientales de la inteligencia
Orígenes de las diferencias de CI entre personas afroestadounidenses y de raza blanca
Cambios de inteligencia con la edad
Cambios generacionales en las calificaciones de pruebas de inteligencia
Resumen
Términos y conceptos clave

Una prueba de inteligencia es una herramienta neutra e inconsecuente hasta que alguien le asigna significado a los resultados obtenidos de ella. Una vez logrado esto, el examinado experimentará muchas sensaciones, que van desde las superficiales hasta aquellas que cambian su vida; serán justas o injustas, útiles o dañinas, apropiadas o engañosas, dependiendo del significado que se vincule con la calificación de la prueba.

Por desgracia, la tendencia a infundirle connotaciones imprecisas e injustificadas a las calificaciones de una prueba de inteligencia es muy generalizada. Es común que legos y estudiantes de psicología se pierdan en errores conceptuales dañinos. A los resultados de una prueba se les asignan valores exagerados o bien se les resta importancia cuando algunas personas los interpretan como predicción de la valía personal, mientras que otros los devalúan por considerarlos triviales e injustos.

El propósito de este tema es aclarar aún más el significado de las calificaciones de una prueba de inteligencia a la luz de la investigación conductual pertinente. En términos específicos, se analizarán cinco aspectos —algunos dirían

que controversias— que tienen que ver con el significado de las calificaciones de una prueba de inteligencia:

- el problema de la parcialidad de prueba.
- los efectos genéticos y ambientales de la inteligencia.
- los orígenes de las diferencias de CI entre afroestadounidenses y caucásicos.
- el destino de la inteligencia en la madurez y vejez.
- los cambios generacionales en las calificaciones de una prueba de inteligencia.

El tema fundamental de esta sección es que las calificaciones de una prueba de inteligencia pueden comprenderse mejor dentro del esquema de la investigación psicológica moderna. Se advierte al lector que las cuestiones de investigación que se tratan aquí son complejas, confusas y, en ocasiones, contradictorias. No obstante, las recompensas de tratar estos temas son sustanciales. Después de todo, el significado de las pruebas de inteligencia se demarca, afina y depura por completo a través de la investigación científica.

LA CUESTIÓN DEL SESGO DE UNA PRUEBA

Sin duda, ninguna práctica dentro de la psicología moderna ha recibido tantos ataques como las pruebas psicológicas. Los críticos le reservan una condena especial, y con frecuencia vehemente, a las pruebas de capacidad. En su amplia respuesta a las miles de críticas dirigidas a las pruebas mentales, Jensen (1980) concluyó que el sesgo de una prueba es el punto en el que suelen concentrarse los críticos. Al declarar el **sesgo de una prueba**, los escépticos expresan de diversas maneras que las pruebas son tan parciales en términos culturales y sexuales que discriminan injustamente a minorías raciales y étnicas, mujeres y pobres. Se citan aquí al pie de la letra una muestra de críticas (Jensen, 1980):

- Las pruebas de inteligencia tienen un nombre tristemente equivocado, debido a que nunca tuvieron el propósito de medir la inteligencia y podrían llamarse de manera más adecuada pruebas de AC (antecedentes culturales).
- Las personas con antecedentes diferentes a los de la cultura en la que se desarrolló la prueba siempre resultarán penalizadas.
- Existen enormes diferencias de clase social en el acceso de un niño a las experiencias necesarias para adquirir las habilidades intelectuales válidas.
- Las calificaciones de CI presentadas para los afroestadounidenses y para los grupos de nivel socioeconómico bajo en EUA reflejan características de la prueba más que de quienes la responden.
- El bajo desempeño de los niños afroestadounidenses en las pruebas convencionales se debe al sesgo en el contenido de las pruebas; es decir, el material de éstas se obtiene de otros sitios ajenos a esa cultura.
- Las mujeres no son tan buenas en matemáticas como los varones porque no han tomado tantas clases de matemáticas en preparatoria y universidad.

¿Son válidas estas críticas? La investigación sobre este asunto resulta ser mucho más complicada de lo que el lector podría suponer. Un aspecto más importante es que las apariencias pueden ser engañosas. Como se explicará más adelante, el hecho de que los reactivos de prueba "parezcan" o "se sientan" como preferenciales para una raza, sexo o clase social no constituye evidencia de sesgo de una prueba. Éste es un asunto objetivo, empírico, no una cuestión de juicio personal.

Aunque los críticos pueden estar reacios a admitirlo, sí existen métodos imparciales y objetivos para investigar el sesgo de una prueba. Uno de los propósitos de esta sección es presentarle dichos métodos al lector. No obstante, una discusión aséptica sobre ecuaciones de regresión y definiciones estadísticas del sesgo de una prueba sería incompleta al dar sólo una mitad de la historia. Los conceptos de sesgo de una prueba están irremediablemente entremezclados con las nociones de equidad de prueba. Una explicación completa de la historia que ro-

dea a la controversia sobre el sesgo de una prueba requiere que se investigue también la cuestión relacionada de la equidad de la misma.

En esta área abundan las diferencias en terminología, por lo cual es importante establecer ciertas distinciones fundamentales antes de proceder. El sesgo de una prueba es un concepto técnico receptivo al análisis imparcial. Más adelante se analizan los métodos más notables para la evaluación objetiva del sesgo de una prueba. En contraste, la equidad de la misma refleja valores sociales y filosofías sobre la utilización de pruebas, en particular cuando ésta se extiende a la selección con el propósito de conceder privilegios u otorgar un empleo. Gran parte de la pasión que rodea a la controversia sobre el sesgo de una prueba se deriva de un fracaso para distinguir dicho sesgo de la equidad de una prueba. Para evitar la confusión, es importante establecer una clara distinción entre estos dos conceptos. Se incluyen análisis separados para el sesgo y la equidad de prueba, comenzando con la pregunta de por qué el sesgo de una prueba es un tema tan polémico.

CONTROVERSIA ACERCA DEL SESGO DE UNA PRUEBA

La controversia acerca del sesgo de una prueba tiene sus orígenes en las diferencias observadas en el CI promedio entre diversos grupos raciales y étnicos. Por ejemplo, los afroestadounidenses obtienen, en promedio, cerca de 15 puntos menos que los estadounidenses de raza blanca en pruebas estandarizadas de CI. Esta diferencia se reduce de 7 a 12 puntos de CI cuando se toman en cuenta las disparidades socioeconómicas. La existencia de estas notables diferencias raciales/étnicas ha alimentado la controversia acerca del sesgo de una prueba. Después de todo, las oportunidades de empleo, admisión a la universidad, obtención del certificado de bachillerato y asignación a clases de educación especial están gobernados, en parte, por los resultados de una prueba. Las pruebas sesgadas pueden perpetuar el legado de la discriminación racial. El sesgo de una prueba es, con absoluto derecho, un tema de intenso escrutinio del público tanto cuanto de las profesiones dedicadas a las pruebas.

Una posibilidad es que las disparidades observadas en el CI indiquen sesgo de una prueba más que diferencias grupales significativas. De hecho, la mayoría de los legos, e incluso algunos psicólogos, considerarían la magnitud de las diferencias raciales en CI como evidencia razonable de que las pruebas de inteligencia tienen un sesgo cultural. Éste es un argumento tentador, pero una gran diferencia entre subpoblaciones definidas no es base suficiente para comprobar la existencia del sesgo en una prueba, sino que debe depender de otros criterios que se establecen más adelante.

Las diferencias raciales y étnicas no son la única fundamentación para la controversia acerca del sesgo de una prueba. También existen diferencias sexuales significativas en algunas medidas de capacidad, en particular en el área del pensamiento espacial (Maccoby y Jacklin, 1974; Halpern, 1986). En un estudio (Gregory, Alley y Morris, 1980), los varones obtuvieron mayores calificaciones que las mujeres en el componente de razonamiento espacial de la *Differential Aptitude Test* que alcanzaron una magnitud de una desviación estándar completa. Dichos datos hacen surgir la posibilidad de que estas pruebas pueden estar sesgadas en favor de los hombres, pero, ¿cómo saberlo?, ¿cuándo las diferencias en las calificaciones de una prueba entre grupos representan sesgo de la misma? Se comenzará con una revisión de los criterios que deberían utilizarse para investigar el sesgo de una prueba de cualquier tipo, ya sea racial, sexual o en relación con cualquier otra característica definitoria.

CRITERIOS DE SESGO Y EQUIDAD DE PRUEBA

El tema del sesgo de una prueba ha recibido amplia atención por parte de psicólogos de la medición, autores de pruebas, periodistas, críticos de pruebas, legisladores y tribunales. Cole y Moss (1989) subrayan una consecuencia

perturbadora de la proliferación de las perspectivas acerca de este tema; es decir, que los conceptos de sesgo de una prueba se han vuelto cada vez más indescifrables y complejos. Lo que es más, la comprensión de éste se vuelve difícil debido a las suposiciones implícitas y con frecuencia emocionales —que sostienen inclusive los académicos— y que pueden conducir a personas honradas a examinar la misma información de diferentes maneras.

En parte, los desacuerdos se perpetúan debido a que los interesados en este debate no pueden aclarar la terminología esencial. Con demasiada frecuencia ocurre que términos como **sesgo de una prueba** y **equidad de prueba** se consideran intercambiables y se usan de manera indefinida, sin definirlos. Se propone que, por lo común, "sesgo de una prueba" y "equidad de prueba" se refieren a aspectos notablemente diferentes del debate acerca de parcialidad de prueba. El examen cuidadoso de ambos conceptos proporcionará una base para un análisis más razonado de este polémico tema.

Según lo interpretan la mayoría de las autoridades en este campo, **sesgo de una prueba** se refiere a índices estadísticos objetivos que examinan el patrón de las calificaciones de una prueba con subpoblaciones apropiadas. Aunque los expertos podrían discrepar acerca de ciertos matices, en la generalidad de los casos existe consenso sobre los criterios estadísticos que indican cuando una prueba está sesgada. Después se hablará más sobre este punto, pero aquí se le proporciona al lector un avance breve: en general, una prueba se considera sesgada si es válida de manera diferencial para diferentes subgrupos; por ejemplo, si las calificaciones de subpoblaciones apropiadas no cayeran dentro de una misma línea de regresión para un criterio pertinente.

En contraste con el estrecho concepto del sesgo de una prueba, la **equidad de una prueba** es una idea amplia que reconoce la importancia de los valores sociales en el uso de pruebas. Incluso si una prueba no está sesgada según los criterios técnicos tradicionales de regresión homogénea, podría considerarse injusta debido a las consecuencias sociales de utilizarla para tomar decisiones de selección. El punto esencial del debate es éste: el sesgo de una prueba (un concepto estadístico) no necesariamente es igual a la equidad de la misma (un concepto de valores). En última instancia, la equidad de prueba se basa en conceptos sociales como la imagen que se tiene acerca de una sociedad justa. En la evaluación de ésta, los valores subjetivos tienen la mayor importancia; los criterios estadísticos de sesgo de una prueba son meramente subordinados. Se regresará a este punto más adelante cuando se analice el vínculo entre valores sociales y equidad de una prueba. Pero se comenzará con una presentación tradicional de los criterios técnicos del sesgo de una prueba.

SIGNIFICADO TÉCNICO DEL SESGO DE UNA PRUEBA: UNA DEFINICIÓN

Una manera útil de analizar el sesgo de una prueba es desde la perspectiva tradicional de la validación. El lector recordará de un capítulo anterior que una prueba es válida cuando una variedad de evidencia sustenta su utilidad y cuando las inferencias obtenidas de ella son apropiadas, significativas y provechosas. Una implicación de este punto de vista es que el sesgo de una prueba puede igualarse con la validez diferencial para diferentes grupos:

> El sesgo está presente cuando la calificación de una prueba tiene significados o implicaciones diferentes para un subgrupo pertinente y definible de individuos evaluados con respecto a los significados o implicaciones para las demás personas examinadas. Así, el sesgo es la validez diferencial de una interpretación dada a la calificación de una prueba para cualquier subgrupo definible y pertinente de individuos examinados (Cole y Moss, 1989).

Quizás un ejemplo concreto ayudará a aclarar esta definición. Supóngase que se utiliza una prueba aritmética de problemas simples expresados con palabras para medir las habilidades de suma de individuos jóvenes. Los problemas podrían expresarse de la siguiente manera: "si tienes dos *six-packs* de refresco, ¿cuántas latas tienes en total?". No obstante, supóngase que

la prueba se utiliza en un grupo de niños de secundaria que proceden de antecedentes de habla hispana. Con éstos, las calificaciones bajas podrían indicar una barrera del lenguaje, no un problema en sus habilidades aritméticas. En contraste, para niños de habla inglesa, las calificaciones bajas indicarían con mayor probabilidad un déficit en habilidades aritméticas. En este ejemplo, la prueba tiene validez diferencial al predecir muy bien las deficiencias aritméticas para niños de habla inglesa, pero de manera inadecuada para los de habla hispana. Según la perspectiva técnica de la validación de una prueba, se concluiría que la prueba está sesgada.

Aunque la definición general de sesgo de una prueba se refiere a la validez diferencial, en la práctica los criterios particulares caen bajo tres encabezados principales: validez de contenido, validez relacionada con el criterio y validez de constructo. Se revisará cada una de estas categorías y se analizarán las conclusiones pertinentes. La cobertura es ilustrativa, no exhaustiva. Se sugiere a los lectores interesados consultar a Jensen (1980), Cole y Moss (1989) y Reynolds y Brown (1984b).

SESGO EN LA VALIDEZ DE CONTENIDO

Es probable que el sesgo en la validez de contenido sea la crítica más común de aquellos que denuncian el empleo de las pruebas estandarizadas con minorías (Hilliard, 1984). Por lo común, los críticos se basan en su propio juicio cuando expresan una o más de las siguientes críticas sobre la validez de contenido de las pruebas de capacidad:

1. Los reactivos piden información que los niños que provienen de minorías o que sufren desventajas no tienen igual oportunidad de aprender.
2. Los reactivos se califican de manera inapropiada debido a que el autor de la prueba ha decidido arbitrariamente cuál es la única respuesta correcta y a los niños de minorías se les califica de modo inadecuado por dar respuestas correctas en su propia cultura, pero no para el autor de la prueba.
3. El fraseo de la pregunta es poco familiar y es posible que un niño que pertenece a una minoría y "sabe" la respuesta correcta no pueda responder porque no comprende la pregunta (Reynolds y Brown, 1984a).

Si cualesquiera de estas críticas fuera precisa, sería una evidencia auténtica de sesgo de una prueba; no obstante, el hecho de expresar una crítica no constituye prueba. El error de estas críticas es que en pocas ocasiones se fundamentan en evidencia empírica.

Reynolds (1982) ha ofrecido una definición de **sesgo del contenido** para pruebas de aptitud que atiende a los puntos anteriores en términos definidos a nivel empírico y que pueden someterse a prueba:

> Un reactivo o subescala de una prueba se considera sesgado en contenido cuando se demuestra que es relativamente más difícil para los miembros de un grupo que para otros, cuando se mantiene constante el nivel de capacidad general de los grupos comparados y no existe ninguna fundamentación teórica razonable que explique las diferencias grupales en el reactivo (o subescala) en cuestión.

Esta definición es útil debido a que propone un enfoque empírico a la cuestión del sesgo de una prueba.

En general, los intentos por probar que los reactivos postulados por los expertos son culturalmente sesgados no han producido la evidencia concluyente que los críticos esperan. McGurk (1953a, 1953b, 1967, 1975) ha escrito de manera extensa sobre este tema y se utilizará su estudio para ilustrar este asunto. Para su tesis doctoral, McGurk le pidió a un panel de 78 jueces (profesores, educadores y estudiantes de posgrado en psicología y sociología) que clasificaran en una de tres categorías cada uno de los 226 reactivos de pruebas estandarizadas de inteligencia ampliamente conocidas. Las categorías eran: menos cultural, neutra, más cultural. McGurk aplicó estos reactivos de prueba a cientos de estudiantes de educación media. Su primer análisis implicó los resultados de prueba de 213 estudiantes afroestadounidenses y 213 estudiantes de raza blanca, equiparados

según currículo, escuela, tiempo de estar inscritos y antecedentes socioeconómicos.

McGurk (1953ab) descubrió que la diferencia media entre estudiantes afroestadounidenses y de raza blanca en el total de la prueba híbrida, expresada en unidades de desviación estándar, era de 0.50. Su comparación de las calificaciones en los 37 reactivos "más culturales" contra los 37 "menos culturales" está más relacionada con el tema del sesgo de una prueba en la validez de contenido. Para los reactivos "más culturales" (aquellos que los jueces consideraron como los de mayor sesgo cultural) la diferencia fue de .30. En el caso de los "menos culturales" (aquellos que los jueces consideraron más justos para los afroestadounidenses y otras minorías culturales) la diferencia fue de .58. En otras palabras, los reactivos designados como más culturales fueron relativamente fáciles para los afroestadounidenses, mientras aquéllos menos culturales fueron relativamente más difíciles. Esta conclusión resultó cierta incluso después de que se tomó en cuenta la dificultad de los reactivos. Además, estas dificultades para los dos grupos se correlacionaron a un nivel casi perfecto ($r = .98$ para los reactivos "más culturales" y $r = .96$ para los "menos culturales"). En esto reside una importante lección que con frecuencia pasan por alto los críticos de las pruebas: los jueces "expertos" no pueden identificar los reactivos sesgados en términos culturales con base en un análisis de las características del reactivo.

En general, con respecto a las pruebas estandarizadas de capacidad y aptitud que más se conocen, la investigación no ha apoyado la creencia popular de que el contenido específico de los reactivos de una prueba es una fuente de sesgo cultural contra las minorías. Esta conclusión no exonera a estas pruebas con respecto a otros criterios de sesgo que se analizan a continuación. Lo que es más, se puede señalar que los autores de una prueba que tengan sentido común deberían estar alertas incluso ante la impresión de sesgo en el contenido de una prueba, dado que la apariencia de inequidad puede afectar de manera bastante tangible las actitudes públicas acerca de las pruebas psicológicas.

SESGO EN LA VALIDEZ PREDICTIVA Y RELACIONADA CON EL CRITERIO

La predicción del desempeño futuro es uno de los usos importantes de las pruebas de inteligencia, capacidad y aptitud. Para esta aplicación en el examen psicológico, la validez predictiva es la forma más importante de validez en relación con el sesgo de una prueba. En general, una prueba imparcial predecirá por igual el desempeño futuro de personas que vienen de subpoblaciones diferentes. Por ejemplo, una prueba imparcial de la aptitud escolar pronosticará el desempeño académico futuro de los estadounidenses de raza negra y blanca con una precisión casi idéntica.

Reynolds (1982) ofrece una definición clara y directa del sesgo de una prueba con respecto a la validez relacionada con el criterio o el **sesgo en la validez predictiva:**

> Una prueba se considera sesgada con respecto a la validez predictiva cuando la inferencia derivada de la calificación de una prueba no se realiza con el menor error aleatorio posible o si existe un error constante en una inferencia o predicción en función de la membresía a un grupo en particular.

Esta definición invoca lo que podría denominarse el criterio de regresión homogénea. Según este punto de vista, una prueba es imparcial si los resultados para todas las subpoblaciones relacionadas se agrupan por igual alrededor de una sola línea de regresión. A fin de aclarar este punto, es necesario introducir conceptos relacionados con la regresión simple. El análisis se basa en el trabajo de Cleary, Humphreys, Kendrick y Wesman (1975).

Supóngase que se está utilizando una prueba de aptitud escolar para predecir el promedio general de calificaciones (pgc) del primer año en la universidad. En el caso de un análisis de regresión simple, la predicción del desempeño futuro se realiza a partir de una ecuación como la siguiente:

$$Y = bX + a$$

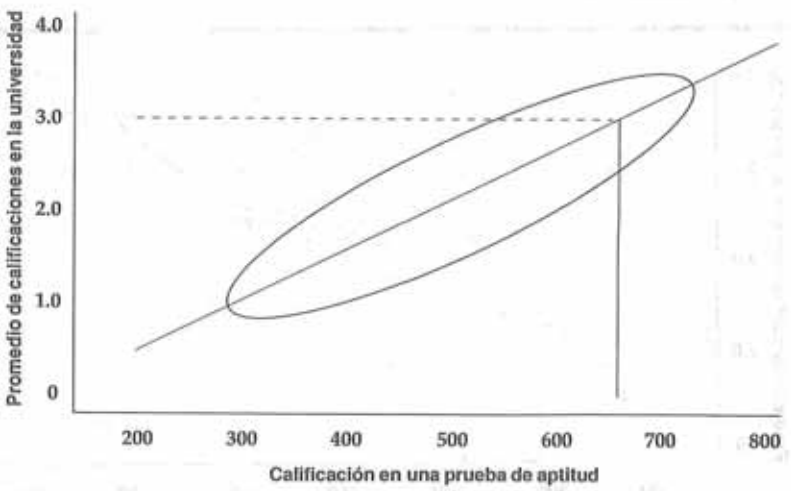
donde Y es el pgc pronosticado para la universidad, X es la calificación en la prueba de

aptitud y *b* y *a* son constantes obtenidas de un análisis estadístico de las calificaciones de una prueba y de estudiantes anteriores. No nos preocuparemos aquí por la manera en que se obtienen *b* y *a*; el lector puede encontrar esta información en cualquier libro sobre estadística elemental.

Los valores de *b* y *a* corresponden a importantes aspectos de la línea de regresión —la línea recta que facilita la predicción más precisa del criterio (calificaciones universitarias) a partir de un medio de predicción (calificación de aptitud; figura 7–4). En particular, *b* corresponde a la pendiente de la línea, donde sus valores más altos indican una pendiente más pronunciada y una predicción más precisa. El valor de *a* representa la intercepción en el eje vertical. Las unidades de medida para *b* y *a* no pueden especificarse de antemano, ya que dependen de las escalas subyacentes utilizadas para *X* y *Y*. Obsérvese en la figura 7–4 que la línea de regresión es la referencia para predecir calificaciones a partir de una calificación de aptitud observada.

De acuerdo con el criterio de regresión homogénea, en una prueba imparcial una sola línea de regresión puede predecir por igual el desempeño de todas las subpoblaciones pertinentes, aunque puedan diferir las medias para los diferentes grupos. Por ejemplo, en la figura 7–5, el grupo A tiene un mejor desempeño que el grupo B, tanto en el medio de predicción como en el criterio. No obstante, la relación entre la calificación de aptitud y las calificaciones es la misma para ambos grupos. En este ejemplo hipotético, la gráfica representa la ausencia de sesgo en la prueba de aptitud con respecto a la validez relacionada con el criterio.

En la figura 7–6 se muestra una situación más compleja que se conoce como sesgo de intercepción. En este caso, las calificaciones para los dos grupos no se agrupan de manera estrecha alrededor de la mejor línea de regresión mostrada como una línea punteada en la gráfica. Se necesitarían líneas de regresión separadas y paralelas (y, por tanto, se requerirían ecuaciones separadas de regresión), para facilitar la predicción precisa. Si se utilizara una sola línea de regresión (la línea punteada), las calificaciones de criterio para el grupo A se sobrepronosticarían, mientras las puntuaciones de criterio para el grupo B se subpronosticarían.



Nota: La línea punteada muestra la manera en que puede emplearse la línea de regresión para predecir el promedio de calificaciones a partir de las calificaciones en una prueba de un solo individuo nuevo.

Figura 7–4. Puntuaciones de prueba, calificaciones y línea de regresión para un gran grupo hipotético de estudiantes universitarios.

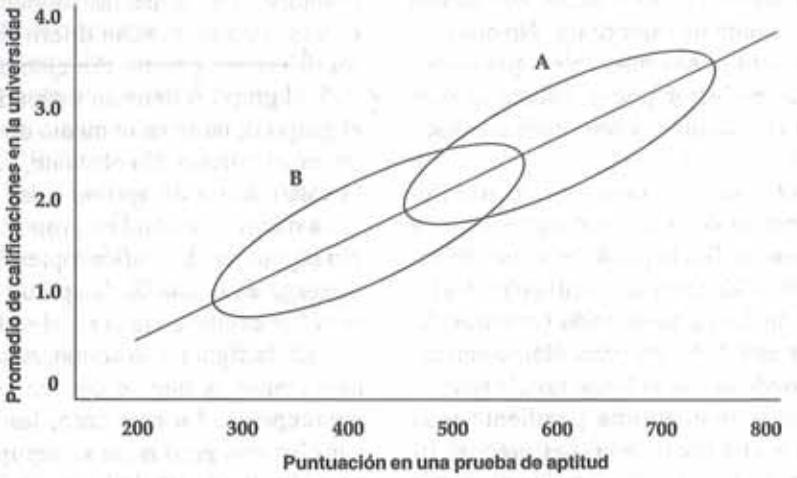


Figura 7–5. Puntuaciones de prueba, calificaciones y línea única de regresión para dos subpoblaciones grandes, hipotéticas, de estudiantes universitarios.

Así, el empleo de una sola línea de regresión constituiría un caso claro de parcialidad de prueba, dado que dicha prueba tiene una validez predictiva diferencial para diferentes subgrupos (contrario a la creencia que se tiene en general, el sesgo de una prueba en estos casos favorece de hecho al grupo con **menor** calificación dado que su desempeño en el criterio se sobrepredice. Por tanto, en ocasiones el sesgo de una prueba puede favorecer a los grupos minoritarios).

Pero, ¿qué sucede si se utilizan líneas de regresión independientes para cada subgrupo? ¿Esto resolvería el problema y rescataría a la

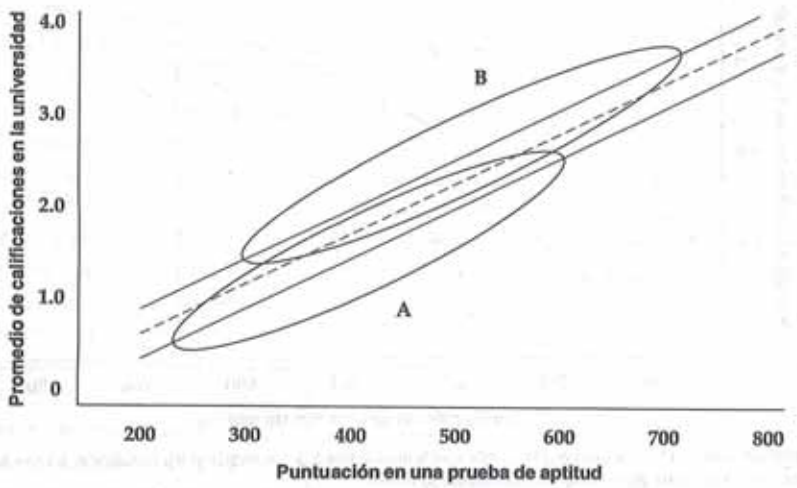


Figura 7–6. Puntuaciones de prueba, calificaciones y líneas paralelas de regresión para dos subpoblaciones grandes, hipotéticas, de estudiantes universitarios.

prueba del sesgo relacionado con el criterio? Las opiniones difieren en cuanto a este punto. Aunque no existe duda de que el uso de ecuaciones de regresión separadas aumentaría la precisión predictiva de la muestra combinada, está a discusión si dicha práctica produciría equidad de la prueba. Se regresará a esta cuestión más adelante, cuando se discuta la pertinencia de los valores sociales en relación con la equidad de la prueba.

Varios investigadores han analizado el sesgo de una prueba en cuanto a la validez relacionada con el criterio de la *Scholastic Aptitude Test* (Prueba de Aptitud Escolar), ahora conocida como *Scholastic Assessment Test* (Prueba de evaluación escolar) que se analiza en un capítulo anterior (Cleary, Humphreys, Kendrick y Wesman, 1975; Breland, 1979; Manning y Jackson, 1984). Una conclusión consistente es que se necesitan líneas independientes y paralelas de regresión para individuos afroestadounidenses y de raza blanca. Por ejemplo, en una escuela, las ecuaciones óptimas de regresión para estudiantes afroestadounidenses, de raza blanca y combinados fue la siguiente:

Afroestadounidenses: $Y = .055 + .0024V + .0025M$

De raza blanca: $Y = .652 + .0026V + .0011M$

Combinados: $Y = .586 + .0027V + .0012M$

donde Y es el pronóstico del promedio de calificaciones universitarias, B es la calificación Verbal de la SAT y M es la calificación de Matemáticas de la SAT (Cleary y col., 1975, p. 29). El efecto de utilizar la fórmula para personas de raza blanca o para la combinación de estudiantes es que se sobrepredicen las calificaciones universitarias de los individuos afroestadounidenses con base en los resultados de la SAT. En la escala tradicional de cuatro puntos; (A = 4, B = 3, etc.), la cantidad promedio de sobrepredicción de 17 estudios independientes fue de .20 o una quinta parte de un punto de calificación (Manning y Jackson, 1984). Lo que significan estos resultados está abierto a debate, pero parece claro que, cuando menos, la SAT y exámenes similares de ingreso no subpredicen las calificaciones universitarias de los miembros de minorías.

En la figura 7–7 se representa el resultado más peculiar de regresión, conocido como sesgo de pendiente. En este caso, las lineas de regresión para subgrupos independientes no son ni siquiera paralelas; por tanto, utilizar una sola de ellas (la línea punteada) para la predicción podría resultar tanto en una subpredicción como en una sobrepredicción de las calificaciones de personas seleccionadas en ambos grupos. La opinión profesional sería unánime en este caso: esta prueba posee alto grado de sesgo en cuanto a la validez relacionada con el criterio.

SESGO EN LA VALIDEZ DE CONSTRUCTO

El lector recordará que la validez de constructo de una prueba psicológica puede documentarse a través de diversas formas de evidencia, incluyendo patrones del desarrollo apropiados en las calificaciones de una prueba; cambios en éstas, debidos a una intervención, que sean consistentes con la teoría; y análisis factoriales confirmatorios. Dado que la validez de constructo es un concepto tan amplio, la definición de **sesgo en la validez de constructo** requiere de un planteamiento general receptivo a la investigación desde una variedad de puntos de vista y con una amplia variedad de métodos. Reynolds (1982) ofrece la siguiente definición:

> Existe sesgo en la validez de constructo cuando se demuestra que una prueba mide diferentes rasgos hipotéticos (constructos psicológicos) en un grupo que en otro o mide el mismo rasgo, pero con diferentes grados de precisión.

Desde una perspectiva práctica, de esta definición provienen dos criterios directos de imparcialidad (Reynolds y Brown, 1984a). Si una prueba es imparcial, entonces las comparaciones entre subpoblaciones pertinentes deberían revelar un alto grado de semejanza en: 1) la estructura factorial de la prueba y 2) la clasificación del nivel de dificultad de los reactivos dentro de la misma. Se analizarán en mayor detalle estos criterios.

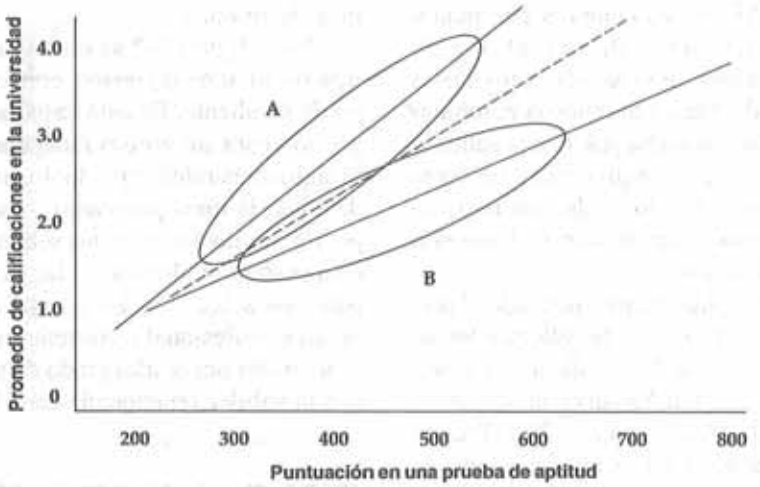


Figura 7–7. Puntuaciones de prueba, calificaciones y líneas no paralelas de regresión para dos subpoblaciones grandes, hipotéticas, de estudiantes universitarios.

Un criterio esencial de la imparcialidad es que la estructura factorial de las calificaciones debe ser invariable entre subpoblaciones apropiadas. Por supuesto, incluso dentro del mismo subgrupo, la estructura factorial de una prueba podría diferir entre grupos por edad, de modo que es importante que se restrinjan las comparaciones entre personas de la misma edad en las subpoblaciones apropiadas. Para individuos de la misma edad, una prueba imparcial poseerá la misma estructura factorial entre subgrupos. En particular, en el caso de una prueba imparcial, el número de factores que surgen y las cargas factoriales de los reactivos o subescalas serán sumamente similares en las subpoblaciones relacionadas.

En general, cuando se realiza de manera independiente el análisis factorial de los reactivos o subescalas de las pruebas importantes de capacidad y aptitud en muestras de raza blanca y minoritarias, surgen los mismos factores en las subpoblaciones de referencia (Reynolds, 1982; Jensen, 1980, 1984). Aunque se han presentado anomalías menores en algunos de estos estudios (Scheuneman, 1987; Reschly, 1978; Gutkin y Reynolds, 1981; Johnston y Bolen, 1984), la investigación en esta área es más notable por sus conclusiones consistentes con respecto a la invarianza factorial entre subgrupos. En el cuadro 7–5 se presenta un resultado completamente común obtenido de la comparación de niños estadounidenses anglosa-jones y de origen mexicano en el WISC-R (Geary y Whitworth, 1988).

Un segundo criterio de imparcialidad en la validez de constructo es que la clasificación de reactivos dentro de una prueba, según el grado de dificultad, debe ser sumamente similar para las subpoblaciones de referencia. Dado que la edad es uno de los principales determinantes de la dificultad del reactivo, esta norma suele verificarse por separado para cada grupo de edad que cubre una prueba. El lector observará lo que este criterio no especifica. No especifica que los subgrupos relacionados deban obtener tasas equivalentes de aprobación en los reactivos de una prueba. Lo que es esencial es que los reactivos que son más (o menos) difíciles para un subgrupo deben ser los más (o menos) difíciles para otras subpoblaciones importantes.

El criterio de la clasificación similar de los reactivos según su dificultad puede someterse a prueba de manera muy directa y objetiva. Si

Cuadro 7–5. Análisis Factorial Comparativo del WISC-R para 100 niños angloestadounidenses y 78 niños méxico-estadounidenses

Subpruebas	Cargas factoriales					
	Comprensión verbal		Organización perceptual		Libertad de la distractibilidad	
	Ang.	M-E	Ang.	M-E	Ang.	M-E
Información	.77	.77				
Semejanzas	.75	.85				
Aritmética					.75	.75
Vocabulario	.86	.86				
Comprensión	.76	.76				
Figuras Incompletas			.58	.58		
Ordenamiento de Dibujos			.77	.61		
Diseño con Cubos			.72	.61		
Ensamble de Objetos			.72	.72		
Claves					.52	.52

Ang. = Angloestadounidenses, N = 100

M-E = México-estadounidenses, N = 78

Fuente: Adaptado con autorización de Geary, D. C. y Whitworth, R. H. (1988). Is the factor structure of the WISC-R different for Anglo- and Mexican-American children? *Journal of Psychoeducational Assessment, 6*, 253-260.

se calcula el nivel de dificultad de cada reactivo por medio del valor *p* (porcentaje de aprobación) para cada subpoblación relacionada, entonces es posible comparar las dificultades relativas de los reactivos entre subgrupos de la misma edad. De hecho, la semejanza en la clasificación de los reactivos según la dificultad para cualquier grupo puede estimarse de manera objetiva por medio de un coeficiente de correlación (r_{xy}). Los valores *p* apareados para los reactivos de una prueba constituyen los valores de *x* y *y* utilizados en el cálculo. Entre más cercano esté el valor de *r* a 1.00, más similar será la clasificación de las dificultades de los reactivos para ambos grupos.

En general, las comparaciones entre grupos en cuanto a la dificultad relativa de los reactivos dentro de las pruebas importantes de capacidad y aptitud han producido correlaciones que se aproximan a 1.00; es decir, la mayor parte de las pruebas muestran clasificaciones sumamente similares de los reactivos según su grado de dificultad entre subpoblaciones importantes (Jensen, 1980; Reynolds, 1982). En un estudio representativo, Miele (1979) investigó la dificultad relativa de los reactivos del WISC con personas afroestadounidenses y de raza blanca en cada uno de los cuatro niveles educativos (preescolar, primero, tercero y quinto grados). Encontró que las correlaciones promedio entre razas (manteniendo constante el nivel de grado) para los valores *p* de los reactivos del WISC fueron de .96 para varones y .95 para mujeres. Estos valores fueron apenas diferentes de las correlaciones entre sexos dentro de la misma raza (manteniendo constante el nivel de grado) donde las correlaciones fueron de .98 (raza blanca) y .97 (afroestadounidenses). Como se señaló, estos datos no son poco comunes. En términos generales, en el caso de las pruebas más conocidas de capacidad y aptitud, la clasificación de los reactivos según el grado de dificultad es prácticamente idéntica para todas las subpoblaciones relacionadas.

REITERACIÓN SOBRE EL SESGO DE UNA PRUEBA

Los críticos que creen que las pruebas están sesgadas en contra de las minorías afirman que las calificaciones de una prueba subestiman la capacidad de los miembros de las minorías. Como se ha analizado en las secciones anteriores, la hipótesis del sesgo de una prueba es una interrogante científica que puede responderse de manera empírica a través de procedimientos como el análisis factorial, las ecuaciones de regresión, las comparaciones intergrupales de los niveles de dificultad para los reactivos "sesgados" e "imparciales" y la clasificación de los reactivos según el grado de dificultad. En general, las pruebas de capacidad y aptitud tienen un desempeño bastante bueno según estos criterios. De hecho, no existe un dominio dentro de las pruebas de capacidad y aptitud donde haya habido evidencia acumulativa que sugiera sesgo de una prueba. Por el contrario, las reseñas amplias de los estudios empíricos rechazan la hipótesis de sesgo (Reynolds, 1982, 1994a; Jensen, 1980; Manning y Jackson, 1984).

Ahora se volverá la atención hacia el concepto más amplio de la equidad de una prueba. ¿Qué tan bien satisfacen los instrumentos existentes los criterios de equidad de una prueba? Como el lector descubrirá, la **equidad de una prueba** implica valores sociales y, por tanto, constituye un tema que es completamente más debatible —y debatido— que el del sesgo.

VALORES SOCIALES Y EQUIDAD DE UNA PRUEBA

Incluso una prueba imparcial podría considerarse injusta debido a las consecuencias sociales de utilizarla en las decisiones sobre selección. En contraste con el concepto limitado y objetivo del sesgo de una prueba, la idea de equidad de una prueba incorpora valores y filosofías sociales del uso de pruebas. Se demostrará al lector que, en última instancia, la aplicación apropiada de las pruebas psicológicas es, en esencia, una conclusión ética que no puede establecerse sólo sobre bases objetivas.

En un artículo que merece mención detallada, Hunter y Schmidt (1976) propusieron la primera distinción clara entre las definiciones estadísticas de sesgo de una prueba y las concepciones sociales de equidad de la misma. Aunque los autores reseñaron con gran precisión los criterios técnicos comunes del sesgo de una prueba, su artículo es más famoso por su descripción de tres posturas éticas incompatibles que pueden y deben afectar el empleo de pruebas.

Hunter y Schmidt (1976) señalaron que las pruebas psicológicas se utilizan con frecuencia para los procedimientos de selección institucional, como el otorgamiento de empleo o la admisión a la universidad. En este contexto, la aplicación de los resultados de una prueba debe guiarse por una filosofía de selección. Por desgracia, en muchas instituciones dicha filosofía es implícita y no explícita. No obstante, cuando se hacen explícitos los valores subyacentes, pueden distinguirse tres posturas éticas: individualismo no calificado, cuotas e individualismo calificado. Dado que estas posiciones éticas se encuentran en el centro de las preocupaciones públicas sobre la equidad de una prueba, se revisarán con cierto detalle.

INDIVIDUALISMO NO CUALIFICADO

En la tradición estadounidense de competencia libre y abierta, la postura del **individualismo no cualificado** dicta que debe seleccionarse, sin excepción, a los candidatos mejor calificados para el empleo, la admisión a una escuela u otros privilegios. Hunter y Schmidt (1976) manifiestan las implicaciones de esta posición:

> Expresado en el lenguaje de los procedimientos de selección institucional, esto significa que una organización debería utilizar cualquier información que posea para hacer una predicción válida científicamente del desempeño de cada individuo y seleccionar siempre a aquellos con el mayor desempeño predicho. Esta posición parece atractiva a primera vista, pero engloba algunas

implicaciones que la mayoría de las personas consideran incómodas. En particular, si la raza, sexo o membresía a un grupo étnico contribuyen a la predicción válida del desempeño en una situación dada, más allá de las contribuciones de las calificaciones de una prueba, entonces aquellos que favorecen el individualismo no calificado están éticamente obligados a utilizar tal medio de predicción.

CUOTAS

La postura ética de las **cuotas** reconoce que muchas instituciones burocráticas y educativas deben gran parte de su existencia a la ciudad o estado en el cual funcionan. Dado que existen por voluntad de las personas, se puede discutir que estas instituciones están éticamente obligadas a actuar de manera "políticamente apropiada" para su localización. La consecuencia lógica de dicha postura son las cuotas. Por ejemplo, en un lugar cuya población sea de un tercio de afroestadounidenses y dos tercios de personas de raza blanca, los procedimientos de selección deben admitir candidatos en aproximadamente la misma proporción. Un procedimiento que se desvía de manera consistente de esta norma se consideraría injusto.

Por definición, las cuotas equitativas se basan, de inicio, en porcentajes de población. Después se tomarían en cuenta los factores que predicen el desempeño futuro, como las calificaciones de una prueba, dentro de las poblaciones relacionadas. No obstante, una consecuencia de las cuotas es que aquellos que se seleccionan no necesariamente tienen las mayores calificaciones en la prueba que se utiliza como medio de predicción.

INDIVIDUALISMO CUALIFICADO

El **individualismo cualificado** es una variación radical del individualismo:

> Esta postura señala que EUA se opone a nivel constitucional a la discriminación con base en la raza, religión, origen nacional o sexo. Una persona que aboga por el individualismo cualificado interpreta esto como un imperativo ético de negarse a utilizar la raza, el sexo y demás como medio de predicción, inclusive si, de hecho, fuera científicamente válido hacerlo (Hunter y Schmidt, 1976).

Con propósitos de selección, el seguidor del individualismo cualificado dependería, exclusivamente, de las capacidades sometidas a prueba, sin referencia a edad, sexo, raza u otras características demográficas. Esto parece loable, pero se analizan las consecuencias potenciales. Supóngase que el individualista cualificado utilizara las calificaciones de la SAT con propósitos de admisión a la universidad. Aunque aquéllas para afroestadounidenses y personas de raza blanca producen líneas independientes de regresión para el criterio de las calificaciones universitarias, el individualista cualificado estaría éticamente obligado a utilizar la línea de regresión menos precisa, obtenida para toda la muestra de solicitantes. En consecuencia, se sobrepredeciría el desempeño futuro de los afroestadounidenses, lo cual en apariencia elevaría la proporción de personas seleccionadas de este grupo de solicitantes. Con respecto a las proporciones de selección, el impacto práctico del individualismo cualificado se encuentra, por tanto, a la mitad entre las cuotas y el individualismo no cualificado.

REITERACIÓN DE LA EQUIDAD DE UNA PRUEBA

¿Cuál filosofía de selección es la correcta? La verdad es que este problema se encuentra más allá del espectro de una solución racional. En un momento u otro, cada una de las posturas éticas expresadas antes han encontrado defensores entre los ciudadanos prudentes, respetados y reflexivos. Sin embargo, no ha surgido un consenso y no es probable que en poco tiempo se llegue a uno. La disputa se reseña aquí:

> ... es típica de los argumentos éticos —la resolución depende en parte de valores irreconciliables. Lo que es más, incluso entre aquellos que concuerdan en sus valores, habrá desacuerdos acerca de la validez de ciertas teorías científicas relacionadas que aún no se han probado de

> manera adecuada. Por tanto, se considera que no existe manera en que pueda resolverse de modo objetivo esta disputa. Cada persona debe elegir lo que bien le parezca (y de hecho, nos encontramos divididos) (Hunter y Schmidt, 1976).

Cuando las posturas éticas chocan —como seguramente ocurre en la aplicación de pruebas psicológicas con propósitos de selección— los sistemas judiciales pueden convertirse en el árbitro final, como se analiza después.

DETERMINANTES GENÉTICOS Y AMBIENTALES DE LA INTELIGENCIA

CONTRIBUCIONES GENÉTICAS A LA INTELIGENCIA

El debate naturaleza-crianza relacionado con la inteligencia es una controversia muy conocida y gastada que en este caso se hará a un lado. Se coincide con MeGue, Bouchard, Iacono y Lykken (1993) en que decenios de estudios sobre adopción, investigación familiar y proyectos con gemelos han probado la existencia de un componente genético sustancial en la inteligencia, aunque los estudios individuales puedan haber sido insuficientes por razones particulares:

> Cuando se toman en conjunto, los estudios de CI con gemelos, familias y adopción, demuestran adecuadamente la existencia de influencias genéticas sobre el CI como cualquiera que pueda lograrse en las ciencias conductuales con métodos no experimentales. Sin postular la existencia de influencias genéticas, simplemente no es posible dar una explicación creíble de la cada vez más consistente semejanza en CI entre gemelos monocigóticos (MC), que entre gemelos dicigóticos (DC) del mismo sexo; la correlación significativa en CI entre familiares biológicos, aunque se les haya criado por separado; y la fuerte relación entre la magnitud de la correlación del CI familiar y el grado de relación genética [p.60].

Por supuesto, la demostración de una influencia genética sustancial en un rasgo no implica que la herencia por sí sola sea responsable de las diferencias entre individuos —los factores ambientales también son formativos, como se revisa más adelante.

La contribución genética a las características humanas como la inteligencia (medida a través de prueba de CI) se evalúa en general en términos de un **índice de hereditabilidad** que puede variar de 0 a 1.0. El índice de hereditabilidad es un estimado de la cantidad de varianza total en un rasgo dado, se debe a factores genéticos. Una hereditabilidad de 0.0 significa que los factores genéticos no contribuyen a la varianza en un rasgo, mientras la hereditabilidad de 1.0 implica que los factores genéticos son los responsables exclusivos de la varianza en un atributo. Por supuesto, para la mayor parte de las características medibles, la hereditabilidad se encuentra entre los dos extremos. MeGue *et al.* (1993) analizan los diversos métodos para calcular la hereditabilidad, con base en los estudios sobre gemelos y adopción.

Es importante destacar que la hereditabilidad es una estadística poblacional que no puede extenderse para explicar la calificación individual. Además, la hereditabilidad de un rasgo dado no es una constante. Como señala Jensen (1969), los estimados de hereditabilidad "son específicos de la población muestreada, el punto en el tiempo, la manera en que se realizaron las mediciones y la prueba particular utilizada para obtener las mediciones". En el caso del CI, la mayor parte de los estudios presentan estimados de hereditabilidad que se encuentran justo alrededor del .50, lo cual significa que cerca de la mitad de la variabilidad en las calificaciones de CI se debe a factores genéticos. No obstante, en algunos estudios la hereditabilidad del CI es mucho más alta, en el rango del .70 (Bouchard, 1994; Bouchard, Lykken, McGue, Segal y Tellegen, 1990; Pedersen, Plomin, Nesselroade y McClearn, 1992).

Podrían citarse docenas de estudios para demostrar la importancia de los factores genéticos en la determinación de la inteligencia como se mide a través de las pruebas de CI; la bibliografía al respecto es casi incontable. Para

una iniciación en el tema, el lector puede consultar a Bouchard (1994) y McGue *et al.* (1993).

Una demostración de lo más fascinante acerca de la contribución genética al CI se encuentra en el *Minnesota Study of Twins Reared Apart* (Estudio Minnesota de gemelos criados por separado; Bouchard *et al.*, 1990). En este estudio en proceso, se reúne a gemelos idénticos criados por separado para aplicarles un extenso examen psicológico. Bouchard (1994) informa que los CI obtenidos de este estudio se correlacionan casi de manera tan elevada como los de gemelos idénticos criados juntos, a pesar de que, con frecuencia, los primeros se vieron expuestos a diferentes condiciones ambientales (en algunos casos, a ambientes que contrastaban en gran medida). En resumen, las diferencias en ambiente parecieron causar muy poca divergencia en los CI de pares de gemelos idénticos criados por separado. Estas conclusiones corroboran una contribución genética a la inteligencia, con una hereditabilidad estimada que se encuentra cerca del .70.

Se puede ilustrar aún más el impulso general de la investigación en esta área con un estudio típico, el nuevo análisis que realizó Honzik (1957) con los datos del estudio sobre adopción de Skodak y Skeels (1949). Este análisis muestra que cuando los niños adoptados se someten varias veces a prueba con un instrumento como la Stanford-Binet, su inteligencia se correlaciona cada vez más con los logros educativos de sus padres biológicos; para los ocho años de edad, la correlación se estabiliza en un valor aproximado de 0.35 (figura 7–8). Es muy interesante que éste es aproximadamente el mismo nivel de correlación que se encuentra entre el CI de los niños y los logros educativos de sus padres en familias intactas. En contraste, Honzik (1957) determinó que la inteligencia de los niños adoptivos se correlacionó en casi cero con los logros educativos de sus padres adoptivos. Los logros educativos de los adultos constituyen una buena representación de su inteligencia —el mejor estimado disponible bajo las circunstancias. Los resultados in-

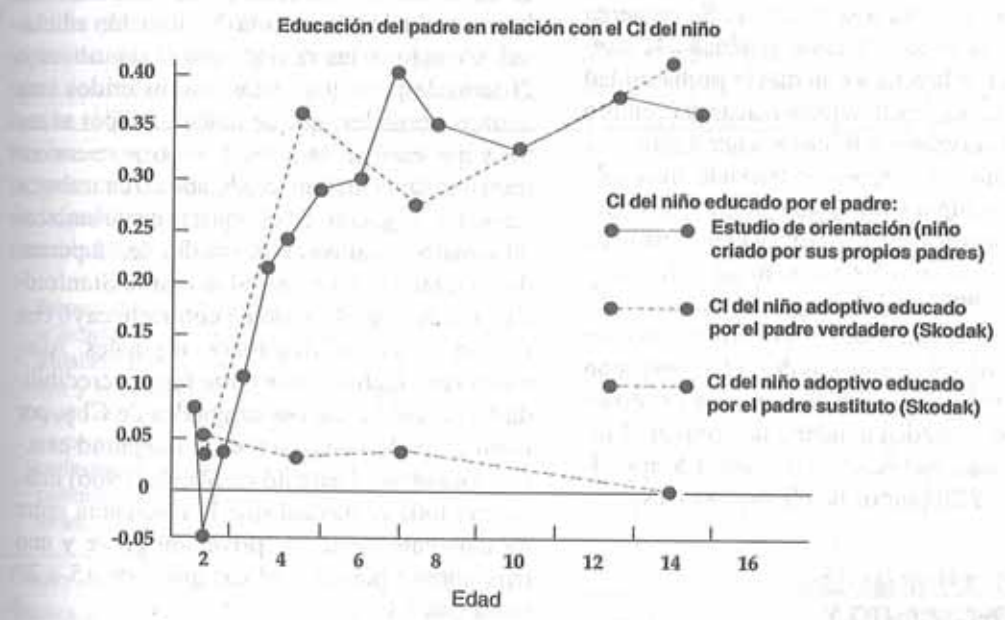


Figura 7–8. Semejanza del CI de niños adoptivos con sus padres sustitutos y biológicos.
Fuente: Reproducido con autorización de: Honzik, M. (1957). Developmental studies of parent-child resemblance in intelligence. *Child Development, 28,* 215-228. Copyright © de The Society for Research in Child Development, Inc.

dicaron que la inteligencia de los hijos adoptivos es análoga a la inteligencia de sus padres biológicos— incluso en su ausencia —pero no muestra relación con la inteligencia de los padres adoptivos— incluso con su presencia constante y cotidiana. ¿Qué significan estas conclusiones? Se sabe que los padres biológicos proporcionaron la contribución genética a la inteligencia de sus hijos, mientras los padres adoptivos aportaron la fundamentación ambiental. El observador desapasionado encontrará difícil evitar una conclusión obvia: los estudios sobre adopción como el presentado por Honzik (1957) corroboran una contribución genética considerable —pero no exclusiva— a la inteligencia.

Sin embargo, se debe evitar la tendencia a considerar cualquier conjunto de investigaciones con una mentalidad simplista de una cosa o la otra. Incluso el teórico más apegado a la explicación hereditaria reconoce que la inteligencia de una persona se moldea también con base en el tipo de experiencia. La pregunta fundamental es: ¿a qué grado es que los ambientes enriquecidos o de privación pueden aumentar o disminuir la inteligencia con respecto al potencial circunscrito a la genética? Se recuerda al lector que la contribución genética a la inteligencia es indirecta, y con mayor probabilidad ocurre a través de estructuras físicas del cerebro y sistema nervioso codificadas a nivel genético. No obstante, el cerebro es bastante maleable ante las manipulaciones ambientales, las cuales pueden aun alterar su peso y la profusión de las redes neuronales (Greenough, Black y Wallace, 1987). ¿Cuánto pueden cambiar dichos impactos ambientales la inteligencia como se mide a través de pruebas de CI? Se revisarán varios estudios que indican que los extremos ambientales pueden ayudar a determinar el resultado intelectual dentro de un rango de aproximadamente 20 puntos de CI y quizá más.

EFECTOS AMBIENTALES: EMPOBRECIMIENTO Y ENRIQUECIMIENTO

En primer lugar, se analizarán los efectos de las desventajas ambientales. Vernon (1979, capítulo 9) ha revisado los primeros estudios sobre privación grave y señala que los niños criados bajo condiciones en las que tuvieron poco o ningún contacto humano pueden mostrar notables mejorías en CI (30 a 50 puntos) cuando se les transfiere a un ambiente más normal. No obstante, se debe ver con escepticismo este conjunto de investigaciones, debido a las condiciones típicamente inusitadas en las que se aplicaron las pruebas iniciales. ¿Se puede aplicar una prueba significativa a niños de siete años criados casi como animales (Koluchova, 1972)?

Típico de estas primeras investigaciones es el estudio de seguimiento de Skeels (1966) con 25 niños huérfanos a los que en un principio se les diagnosticó retraso mental (Skeels y Dye, 1939). Estos niños se evaluaron inicialmente cuando tenían cerca de un año y medio de edad mientras vivían en un orfanatorio en el que se les daba muy poca estimulación. A 13 de ellos se les transfirió a otro hogar, donde recibieron atención supervisada y afectuosa de parte de niñas con retraso mental de mayor edad. Estos niños mostraron un aumento considerable en CI, mientras los 12 que permanecieron en el lugar original tuvieron una disminución adicional. Cuando se les rastreó para el seguimiento 26 años después, los 13 casos transferidos eran adultos normales, que se mantenían por sí solos y que estaban casados. Los otros se encontraban todavía institucionalizados o en trabajos humildes. El grupo con enriquecimiento ambiental mostró un aumento promedio de 32 puntos de CI cuando se les reevaluó con la Stanford-Binet, mientras el grupo de contraste cayó con respecto a sus calificaciones originales. Aunque no nos inclinamos a darle mucha credibilidad a las calificaciones originales de CI y, por tanto, se podría discrepar de la magnitud exacta del cambio, el estudio de Skeels (1966) indica con toda seguridad que la diferencia entre un ambiente inicial de privación grave y uno más normal podría explicar quizá de 15 a 20 puntos de CI.

Jensen (1977) presenta una demostración metodológicamente novedosa de un déficit en el CI inducido por el ambiente. Este investigador examinó a 653 niños de raza blanca y a 826

niños afroestadounidenses de una pequeña población rural en la parte sudoriental de Georgia. La hipótesis alterna de este estudio era que los niños afroestadounidenses de mayor edad tendrían calificaciones más bajas que sus hermanos menores, debido a los efectos acumulativos y reductores del intelecto de su ambiente inhóspito y de profunda privación. Según la hipótesis del déficit acumulativo, los efectos acumulativos de las desventajas ambientales en los factores relacionados con el desarrollo mental dan por resultado una tendencia a la disminución consistente del CI. En contraste, los niños de raza blanca, que tienen menos privaciones ambientales, no deberían mostrar un déficit acumulativo intelectual como una función lineal de la edad.

A todos se les aplicó la *California Test of Mental Maturity* (Prueba California de Madurez Mental; CTMM, revisión de 1963), una prueba estandarizada de la inteligencia general, como parte de un programa estatal obligatorio de evaluación. La CTMM produjo calificaciones de CI de desviación estandarizadas de manera cuidadosa (media nacional de 100, desviación estándar de 15) que se calcularon por separado de las normas nacionales para cada nivel desde el jardín de niños hasta bachillerato. Jensen (1977) señaló que las poblaciones muestreadas, en particular el grupo afroestadounidense, no tenían el propósito de ser representativas de la población general de EUA, ya fuera de raza blanca o de afroestadounidenses:

> Es probable que las personas de raza negra en la localidad bajo estudio se encuentren en una situación de desventaja tan grave, a nivel educativo y económico, como la que pudiera encontrarse en cualquier parte de EUA. Si en este grupo no existe una disminución con la edad, parecería muy dudoso que pudiera encontrarse en cualquier subpoblación dentro de nuestras fronteras.

Como se pronosticó, los niños afroestadounidenses de mayor edad obtuvieron calificaciones inferiores a las de sus hermanos menores y la magnitud de la diferencia se relacionó de manera directa con la diferencia en edad. En particular, los niños afroestadounidenses parecieron perder, en promedio, aproximadamente un punto de CI por año entre los 6 y 16 años de edad, con una pérdida acumulativa total de 5 a 10 puntos de CI. La cantidad exacta de la pérdida depende de la manera en que se interpreten algunas peculiaridades aparentes de la muestra en los datos (figura 7–9). Además, si se incluye en el análisis el probable déficit de CI que haya ocurrido entre el nacimiento y los cinco años de edad, se puede conjeturar que el efecto total de la privación ambiental es sustancialmente mayor que la disminución de 5 a 10 puntos de CI que informó Jensen (1977).

Scarr y Weinberg (1976, 1983) revirtieron la pregunta realizada por Jensen (1977): ¿qué sucede con la inteligencia de los niños afroestadounidenses cuando familias de raza blanca que poseen mayores recursos económicos y educativos los adoptan dentro de un ambiente relativamente enriquecido? Como se analiza después, es bien conocido que los niños afroestadounidenses criados por sus propias familias obtienen calificaciones de CI que promedian cerca de 15 puntos por debajo de los CI de los de raza blanca (Jensen, 1980). Es probable que una parte de esta diferencia —y quizá toda— se deba a las muchas diferencias sociales, económicas y culturales entre estos dos grupos. Por el momento, se hace a un lado este asunto y en su lugar se atenderá a una pregunta que tiene



Figura 7–9. Diferencia promedio en el CI en la CTMM entre hermanos afroestadounidenses menores y mayores en función de la diferencia en edades.
Fuente: Basado en datos de Jensen, A.R. (1977). Cumulative deficit in IQ of blacks in the rural south. *Developmental Psychology, 13*, 184-191.

que ver con la maleabilidad del CI: ¿se logra alguna diferencia cuando se adopta a los niños afroestadounidenses dentro de un ambiente con mayores ventajas económicas y educativas?

Scarr y Weinberg (1976, 1983) encontraron que 130 niños afroestadounidenses e interraciales, adoptados por familias de clase media alta, promediaron un CI Total de 106 puntos en la Stanford-Binet o el WISC, seis puntos por encima del promedio nacional, y entre 18 y 21 puntos por arriba de las calificaciones que por lo común se encuentran con individuos afroestadounidenses. Los niños afroestadounidenses adoptados antes del año de edad tuvieron resultados que incluso fueron mejores, con un CI promedio de 110 puntos. Sólo podría uno preguntarse cuáles habrían sido las calificaciones de CI si las adopciones hubieran ocurrido al nacer y si se hubiera proporcionado excelente cuidado prenatal. Este estudio indica que cuando el ambiente inicial es óptimo, el CI puede elevarse en, tal vez, 20 puntos.

Además de enriquecer el ambiente psicológico, otro enfoque para incrementar el CI ha consistido en mejorar el nivel nutricional de los niños. Un estudio de Schoenthaler, Amos, Eysenck y colaboradores (1991) examinó los efectos de los complementos de vitaminas-minerales sobre el CI de 615 estudiantes de educación media. Excepto por el hecho de que se utilizó un diseño experimental extradamente cuidadoso con asignación aleatoria doble ciego, sería fácil descartar este estudio como una investigación "charlatana" sobre la improbable hipótesis de que 12 semanas de enriquecimiento con vitaminas y minerales elevaría el CI. Los investigadores encontraron que, en comparación con un grupo control que recibió un placebo, los niños que recibieron 100% de los complementos diarios recomendados de vitaminas/minerales no mostraron aumento en el CI Verbal, pero sí una elevación de 3.7 puntos en el CI de Ejecución del WISC-R. Este resultado requiere que otros investigadores lo repliquen y extiendan. Lynn (1993) revisa estudios adicionales, incluyendo una ganacia informada de nueve puntos de CI en razonamiento no verbal entre niños que recibieron un complemento de vitaminas-minerales durante un periodo de ocho meses.

Las limitaciones de espacio impiden una discusión más detallada de los efectos ambientales sobre el CI. No obstante, es necesario mencionar que ha surgido gran cantidad de bibliografía a partir de los primeros estudios de intervención y enriquecimiento-estimulación con niños que están en riesgo de fracaso escolar y retraso mental (p. ej., Ramey y Ramey, 1998; Spitz, 1986; Sprigle y Schaeffer, 1985). En general, estos estudios muestran que la intervención y el enriquecimiento pueden elevar el CI de niños que están en riesgo de fracasar en la escuela o sufrir retraso mental. En un resumen de cuatro decenios de investigación, Ramey y Ramey (1998) extrajeron seis principios de la investigación sobre intervención temprana con niños en situación de riesgo. Estos autores los definen como "consistencias notables en las principales conclusiones" de los estudios de investigación:

1. Las intervenciones que comienzan temprano (p. ej., durante la lactancia) y continúan por más tiempo, proporcionan los mejores beneficios a los niños participantes.
2. Las intervenciones más intensivas (p. ej., número de visitas por semana) producen efectos positivos más amplios que las menos intensivas.
3. Las experiencias directas de enriquecimiento (p. ej., el trabajo directo con los niños) proporcionan mayor impacto que las experiencias indirectas.
4. Los programas con servicios completos (p. ej., mejorías múltiples) producen mayores cambios positivos que aquellos con un enfoque estrecho.
5. Algunos niños (p. ej., aquellos con peso normal al nacer) muestran mayores beneficios de la participación que otros.
6. Los beneficios iniciales positivos disminuyen a lo largo del tiempo si el ambiente del niño no estimula las actitudes positivas y el aprendizaje continuo.

Una preocupación acerca de los programas de intervención temprana es su costo, que ha sido

excesivo para algunos proyectos de demostración. Los escépticos tienen dudas acerca de los aspectos prácticos y, también, sobre la ganancia de proporcionar intervención continua, extensa y con una base amplia, casi desde el nacimiento para millones de niños que están en riesgo de presentar problemas del desarrollo. Ésta es una preocupación realista porque "relativamente pocos programas de intervención temprana han tenido un seguimiento a largo plazo" (Ramey y Ramey, 1998). Los críticos también se preguntan si los programas tan sólo le enseñan a los niños la manera en que se realiza una prueba sin afectar mucho su inteligencia subyacente (Jensen, 1981). Por último, está la cuestión de la congruencia cultural. Los programas de intervención han sido diseñados principalmente por psicólogos de raza blanca y después se aplican de manera desproporcionada a niños de minorías. Ésta es una preocupación debido a que los programas deben ser culturalmente pertinentes y aceptables para los consumidores, de otro modo las intervenciones están condenadas al fracaso.

EFECTOS TERATOGÉNICOS SOBRE LA INTELIGENCIA Y EL DESARROLLO

En el desarrollo normal prenatal, el feto se encuentra protegido del ambiente exterior por la placenta, que es un órgano vascular dentro del útero a través del cual se nutre al embrión. No obstante, algunas sustancias conocidas como **teratógenos** atraviesan la barrera placentaria y provocan deformidades físicas en el feto. Si éstas se presentan en el cerebro, principalmente, pueden producir trastornos conductuales durante el resto de la vida, que incluyen bajo CI y retraso mental. La lista de teratógenos potenciales es casi infinita e incluye medicamentos, hormonas, drogas ilícitas, tabaco, alcohol, radiación, químicos tóxicos e infecciones vitales (Berk, 1989; Martin, 1994). Se destacará de manera breve al más frecuente de los teratógenos, el alcohol, que también es el más fácil de prevenir.

El consumo excesivo de alcohol en mujeres embarazadas causa que sus hijos estén en muy alto riesgo de sufrir el **síndrome alcohólico fetal** (SAF), que es un conjunto específico de anormalidades que describieron inicialmente Jones, Smith, Ulleland y Streissguth (1973). Los niños con SAF tienen una inteligencia notablemente menor; cuando se les evalúa en la adolescencia o la edad adulta, cerca de la mitad se encuentra dentro del rango de retraso mental en pruebas de CI (Olson, 1994). La exposición prenatal al alcohol es una de las principales causas de retraso mental que se conocen en el mundo occidental. Los criterios que definen al SAF incluyen:

1. Retraso en el crecimiento prenatal, posnatal o ambos —peso menor al percentil 10 después de hacer la corrección para edad de gestación.
2. Disfunción del sistema nervioso central —deformaciones de cráneo o cerebro, retraso mental de leve a moderado, anormalidades neurológicas y problemas de conducta.
3. Dismorfología facial —espacio amplio entre los ojos, pequeña apertura entre los párpados, nariz pequeña y respingada, labio superior delgado y deformidades menores en las orejas (Clarren y Smith, 1978; Sokol y Clarren, 1989).

El SAF completo que se describe antes ocurre principalmente en los hijos de mujeres alcohólicas —aquellas que consumen muchas bebidas en cada ocasión. Con menores niveles de ingestión de alcohol puede surgir una manifestación más leve del síndrome conocida como **efecto alcohólico fetal**. Por lo común, un niño con este efecto tiene una apariencia física normal, pero exhibe problemas evidentes en su capacidad de atención y es más lento para responder en un paradigma de tiempo de reacción (Streissguth, Martin, Barr y Sandman, 1984). Además, el efecto es lineal en relación con la dosis; es decir, puede ser que no exista un nivel seguro de ingestión de alcohol durante el embarazo (Streissguth, Bookstein y Barr, 1996). Por tal razón, en la actualidad los médicos suelen recomendar la abstención del alcohol durante el embarazo.

No obstante, un **conservador** estimado de la incidencia del SAF (formas leves a graves)

en el mundo occidental es de 1 por cada 1 000 nacimientos vivos, donde la mayor parte de los casos quedan sin diagnosticar o reconocerse (Abel, 1995). Spohr y Steinhausern (1996) proporcionan una excelente reseña de la investigación acerca del SAF.

EFECTOS DE LAS TOXINAS AMBIENTALES SOBRE LA INTELIGENCIA

Muchos químicos y subproductos industriales pueden alterar temporalmente el sistema nervioso o incluso causar daños permanentes que afectan la inteligencia. Algunos ejemplos incluyen el plomo, mercurio, manganeso, arsénico, talio, tetraetilo de plomo, compuestos orgánicos de mercurio, bromuro de metilo y bisulfuro de carbono (Lishman, 1978, capítulo 13). Con seguridad, la más estudiada de estas toxinas ambientales es el plomo, el cual se analiza aquí brevemente.

Las fuentes de absorción de plomo en humanos incluyen: ingestión de pintura con pigmentos de plomo en lactantes e infantes; aspiración de partículas de plomo de las emisiones de fundiciones o de la combustión de gasolinas con plomo en los automóviles; ingestión de comida de latas con soldadura de plomo o de cerámica con glaseado de plomo e ingestión de agua que ha pasado por tuberías de plomo. Dado que el cuerpo humano secreta el plomo de manera lenta, la mayoría de las personas en países industrializados llevan una carga de este metal sustancialmente mayor —quizá 500 veces más alta— que la conocida en tiempos anteriores a Roma (Patterson, 1980).

Cualquier investigador médico y psicológico que haya estudiado este tema reconoce los peligros del alto nivel de exposición al plomo; las altas dosis se vinculan de manera irrefutable con la parálisis cerebral, trastornos epilépticos, ceguera, retraso mental e incluso la muerte. La pregunta más importante tiene que ver con la exposición "asintomática" al plomo: ¿un nivel de absorción insuficiente para provocar síntomas médicos evidentes puede producir de todas maneras una disminución en las capacidades intelectuales?

Las conclusiones de la investigación sobre el tema son complejas y polémicas. Al utilizar como índice de acumulación el plomo encontrado en los dientes de leche mudados por niños pequeños, Needleman, Gunnoe, Leviton, Reed, Peresie, Maher y Barrett (1979) informaron que la exposición asintomática al plomo se relacionaba con disminuciones en la inteligencia general (cerca de 4 puntos de CI) y con un desempeño menor en subpruebas verbales, pruebas de procesamiento auditivo y del habla y en una medida de atención con base en el tiempo de reacción. Estas diferencias persistieron 11 años después (Needleman, Schell, Bellinger, Leviton y Allred, 1990). No obstante, al utilizar un método de estudio similar, Smith, Delves, Lansdown, Clayton y Graham (1983) encontraron un efecto no significativo de la exposición de los niños al plomo cuando se controlaron los factores de nivel educativo de los padres y nivel social.

Las conclusiones de la investigación sobre este tema son contradictorias, en parte, debido a que es difícil separar los efectos del plomo de aquéllos de la pobreza, estrés, nutrición deficiente y otras variables que producen confusión (Gregory y Mohan, 1977). Es más probable que la exposición asintomática al plomo tenga efectos dañinos sobre el sistema nervioso central que se traduzcan en reducción de la inteligencia, dificultades de atención y en una multitud de consecuencias conductuales indeseables. Aunque los científicos no coincidan sobre este punto, se recomienda reducir al menor nivel posible la exposición al plomo en humanos.

ORÍGENES DE LAS DIFERENCIAS DE CI ENTRE PERSONAS AFROESTADOUNIDENSES Y DE RAZA BLANCA

DIFERENCIAS DE CI ENTRE PERSONAS AFROESTADOUNIDENSES Y DE RAZA BLANCA

Como se señaló antes, los afroestadounidenses obtienen, en promedio, cerca de 15 puntos me-

nos en pruebas de CI que los estadounidenses de raza blanca. La diferencia no es trivial, ya que casi alcanza una desviación estándar completa. Aunque la diferencia en CI fluctúa de un análisis a otro —tan pequeña como 10 puntos en algunos estudios, pero tan grande como 20 puntos en otros—, la disparidad se ha documentado con numerosas muestras y con una amplia variedad de pruebas. El control de la variable de clase social reduce la diferencia, pero no la elimina. Lo que es más, la magnitud de la diferencia parece haber permanecido inmune al cambio desde mediados hasta finales del siglo XX (Jensen, 1980; Kennedy, Van de Riet y White, 1963; Shuey, 1966).

La figura 7–10 presenta uno de los primeros estudios sobre las diferencias entre personas afroestadounidenses y de raza blanca que se basó en la edición de 1960 de la Stanford-Binet. El lector observará que, en promedio, la muestra de raza blanca ($M = 101.8$) tuvo calificaciones más altas que la muestra afroestadounidense ($M = 80.7$) en cerca de 20 puntos de CI. Recientemente ha surgido un patrón similar con el WAIS-R, donde la muestra de estandarización de individuos de raza blanca ($M = 101.4$) tuvo calificaciones mayores en 14.5 puntos de CI que la muestra de personas afroestadounidenses ($M = 86.9$) (Reynolds, Chastain, Kaufman y McLean, 1987). La diferencia racial también surgió en Stanford-Binet: Cuarta Edición actualizada, donde los individuos de raza blanca ($M = 103.5$) dentro de la muestra de estandarización obtuvieron calificaciones más altas que los afroestadounidenses ($M = 86.1$) en casi 17.5 puntos (Thorndike, Hagen y Sattler, 1986).

Cuando se toman en cuenta variables demográficas como el nivel socioeconómico, el tamaño de la diferencia promedio se reduce de .5 a .7 desviaciones estándar o de 7 a 10 puntos de CI pero permanece constante (Reynolds y Brown, 1984a). La veracidad de una diferencia racial en calificaciones de CI se ha replicado

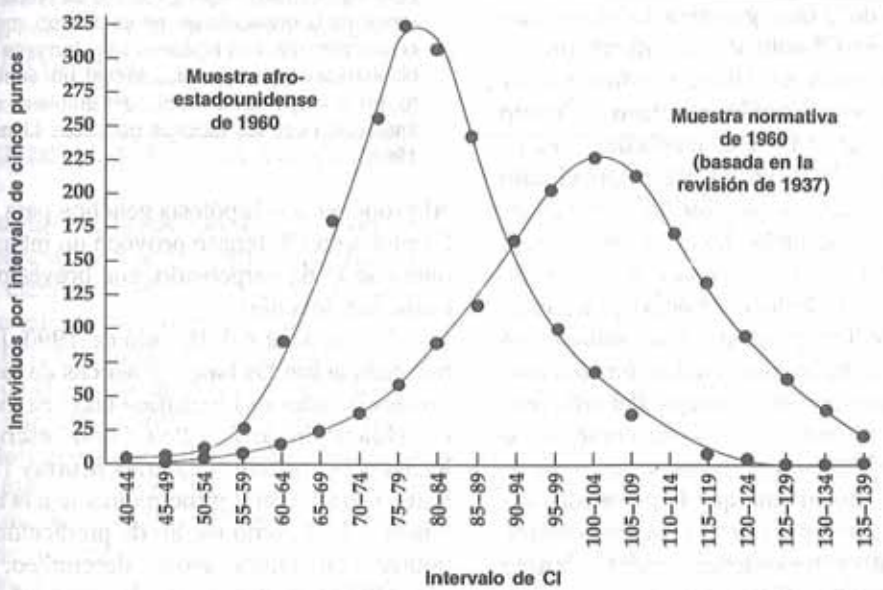


Figura 7–10. Distribuciones de CI de la Stanford-Binet para estadounidenses de raza negra y blanca. **Fuente:** Reproducido con autorización de Kennedy, W. A. Van de Riet, V. y White, Jr., J. C. (1963). A normative sample of intelligence and achievement of negro elementary school children in the southeast United States. *Monographs of the Society for Research in Child Development, 28* [No. 90], 68. Copyright © de la Society for Research in Child Development, Inc.

en tantas ocasiones que ya no constituye un punto de disputa seria. No obstante, la **interpretación** de estas conclusiones bien documentadas se debate con vehemencia.

Un punto de vista (que se analizó antes) es que la disparidad observada en el CI es producida, en parte o totalmente, por el sesgo de una prueba. Es una perspectiva que se sostiene de manera amplia a nivel popular y que, como se vio antes, en pocas ocasiones tiene apoyo de los estudios técnicos sobre el sesgo de una prueba. Éste puede representar una función menor en las diferencias raciales, pero no puede explicar las diferencias grandes y pesistentes en calificaciones de CI entre estadounidenses de raza blanca y de origen africano. En este caso se intenta examinar una hipótesis diferente: ¿la diferencia entre estadounidenses de raza blanca y de origen africano es principalmente genética?

HIPÓTESIS GENÉTICA DE LAS DIFERENCIAS RACIALES EN CI

La hipótesis de la base genética de las diferencias raciales en CI obtuvo inicialmente importancia académica en 1969, cuando Arthur Jensen publicó un artículo intitulado "¿Cuánto podemos elevar el CI y el rendimiento escolar?" (Jensen, 1969). Jensen estableció el tono de su artículo en la oración introductoria cuando afirmó que "se ha hecho el intento de proporcionar educación compensatoria y en apariencia esto ha fallado". Además, este autor expresó que los programas de educación compensatoria se basaban en dos fundamentos teóricos engañosos: el "concepto del niño promedio", que considera a los niños como más o menos homogéneos, y la "hipótesis de la privación social", que afirma que la privación ambiental es la principal causa de la disminución en rendimiento y en calificaciones de CI. Jensen adoptó una postura firme en contra de ambas suposiciones. Lo que es más, basándose principalmente en la bibliografía sobre genética conductual, Jensen implicó que la razón por la que los individuos de raza blanca tenían calificaciones más altas que los afroestadounidenses en pruebas de CI estaba relacionada probablemente más con factores genéticos que con los efectos de la privación ambiental. La idea general de su artículo era la de sugerir que, dado que la educación compensatoria ha resultado ineficiente y debido a que la evidencia sugiere que existe un fuerte componente genético relacionado con el CI, sería apropiado considerar una explicación genética para la diferencia bien documentada en favor de los individuos de raza blanca en las pruebas de CI. Jensen formuló la hipótesis genética de una manera cuidadosa, tentativa y académica:

> El hecho de que una hipótesis razonable no se haya probado de manera rigurosa no significa que se le debería rechazar de inmediato; sólo que se necesita investigación más apropiada para ponerla a prueba. Considero que tal investigación definitiva es totalmente posible, pero no se ha hecho. De modo que nos quedan varias líneas de evidencia, ninguna de las cuales es definitiva por sí sola, pero que, vistas en conjunto, hacen que sea razonable la hipótesis de que los factores genéticos tienen una fuerte participación en la diferencia promedio en inteligencia entre personas de raza negra y blanca. La preponderancia de la evidencia es, en mi opinión, menos consistente con una hipótesis estrictamente ambiental que con una genética, lo cual, por supuesto, no excluye la influencia del ambiente o su interacción con los factores genéticos (Jensen, 1969).

Al proponer una hipótesis genética para las diferencias en CI, Jensen provocó un intenso debate que se ha perpetuado, con breves pausas, hasta nuestros días.

A mediados del decenio de 1990, la controversia sobre las bases genéticas de las diferencias raciales se intensificó una vez más con la publicación de *The Bell Curve*, escrito por Richard Herrnstein y Charles Murray (1994). Este libro se refería principalmente a la importancia del CI como medio de predicción de la pobreza, abandono escolar, desempleo, paternidad ilegítima, delincuencia y otras patologías sociales. Pero dos capítulos sobre las diferencias étnicas en inteligencia provocaron una protesta general entre los científicos sociales y el público lego. Los autores revisaron docenas de estudios y concluyeron que la brecha en CI

entre personas afroestadounidenses y de raza blanca había cambiado poco en este siglo. También discutieron que el sesgo de una prueba no podía explicar las diferencias raciales. Lo que es más, señalaron que las razas no sólo diferían en las calificaciones promedio de CI, sino también en el perfil de capacidades intelectuales. Además, concluyeron que la inteligencia sólo puede modificarse de manera limitada, incluso con una intervención ambiental intensiva. Al igual que Jensen, Herrnstein y Murray (1994) expresaron su hipótesis genética con mucho cuidado:

> Nos parece sumamente probable que tanto la genética como el ambiente tienen algo que ver con las diferencias raciales [en capacidad cognitiva]. ¿Cuál podría ser la composición? Nos mostramos resueltamente agnósticos al respecto: hasta donde podemos determinarlo, la evidencia no justifica todavía un estimado.

Aunque los autores declinaron proporcionar un estimado de la contribución genética a las diferencias raciales en CI, es claro, por el tono pesimista de su libro, que lo consideran sustancial. ¿Tal conclusión tiene fundamentación con base en la evidencia?

VALIDEZ DE LA HIPÓTESIS GENÉTICA

La hipótesis genética de las diferencias raciales en CI es una idea poco común que muchos legos y científicos sociales consideran anatema, pero esto no significa ninguna objeción, y la superficialidad no sustituye a un examen razonado de la evidencia. Ante los resultados de análisis e investigación adicionales, ¿es válida la hipótesis genética para las diferencias en CI? Aquí se examinarán tres líneas de evidencia que indican que la respuesta es "no".

Varios críticos han señalado que la hipótesis genética se basa en la suposición cuestionable de que la evidencia de la hereditabilidad del CI dentro de los grupos puede utilizarse para inferir la hereditabilidad entre grupos raciales. Jensen (1969) expresó esta premisa de manera bastante explícita al destacar que el componente genético sustancial en el CI es una evidencia sugerente de que las diferencias en éste entre estadounidenses de raza blanca y de origen africano tienen, en parte, una base genética. Haciendo eco de críticos anteriores, Kaufman (1990) responde como sigue:

> No puede inferirse la hereditabilidad entre grupos a partir de estudios que hayan proporcionado evidencia de la hereditabilidad del CI dentro de éstos. Incluso si el CI se heredara por igual dentro de las razas negra y blanca por separado, eso no prueba que las diferencias en CI entre las razas tengan un origen genético. El sencillo ejemplo de Scarr-Salapatek (1971, p. 1226) explica muy bien este punto: plante en dos tipos de suelo —condiciones buenas contra condiciones pobres— dos muestras de semillas elegidas de manera aleatoria de una población genéticamente heterogénea y compare la altura de las plantas completamente crecidas. Dentro de cada tipo de suelo, las variaciones individuales en altura se determinan de manera genética; pero la diferencia promedio en altura entre las dos muestras está sólo en función del ambiente.

Un segundo problema con la hipótesis genética es que una predicción importante derivada de ella no se sostiene ante el examen empírico. En términos específicos, si las diferencias raciales en CI tienen una base genética, entonces el grado de ascendencia africana debería ayudar a pronosticar los CI promedio dentro de los subgrupos afroestadounidenses. En otras palabras, los afroestadounidenses con ascendencia totalmente africana deberían tener, en promedio, calificaciones menores en pruebas de CI que los afroestadounidenses con una mezcla de ascendencia africana y caucásica. No obstante, al utilizar marcadores sanguíneos de grupo como índice de la mezcla negra-caucásica, Scarr, Pakstis, Katz y Barker (1977) no encontraron evidencia de una relación entre el grado de ascendencia blanca y las habilidades intelectuales dentro de la población afroestadounidense. Así, sus resultados no sustentaron una fuerte hipótesis de diferencias genéticas raciales en inteligencia.

Otra crítica de la hipótesis genética es que el análisis cuidadoso de los factores ambientales proporciona una explicación suficiente de las

diferencias raciales en CI; es decir, la hipótesis genética es simplemente innecesaria. Éste es el enfoque que toman Brooks-Gunn, Klebanov y Duncan (1996) en un estudio con 483 niños afroestadounidenses y de raza blanca con bajo peso al nacer. Lo que hace diferente a su estudio de otros análisis similares es la riqueza de sus datos. En vez de utilizar sólo 1 o 2 medidas del ambiente (p. ej., un solo índice del nivel de pobreza), obtuvieron datos longitudinales sobre nivel de ingresos y muchos otros factores relacionados con la pobreza, como el tiempo que se permanece en el hospital, capacidad verbal materna, ambiente de aprendizaje en el hogar, condiciones del vecindario y otros componentes de la clase social de la familia. Cuando se evaluó el CI de los niños a los cinco años de edad utilizando el WPPSI, los investigadores encontraron la frecuente disparidad entre los niños de raza blanca (CI medio de 103) y los afroestadounidenses (CI medio de 85). No obstante, cuando se controlaron la pobreza y sus factores relacionados, las diferencias de CI casi se eliminaron por completo. Este estudio sugiere que las investigaciones anteriores habían subestimado los profundos efectos de la pobreza y de sus factores relacionados como contribución a las diferencias de CI entre personas afroestadounidenses y de raza blanca. Neisser, Boodoo, Bouchard y colaboradores (1996) ofrecen perspectivas adicionales sobre las diferencias raciales en CI y temas relacionados.

Antes de abandonar el tema de las diferencias raciales en CI, se debe señalar que la percepción vinculada con este tema es, en gran medida, inmerecida por dos razones. Primero, los grupos raciales siempre han mostrado valores de CI semejantes —lo cual significa que los pueblos de la tierra son mucho más parecidos que diferentes. Segundo, como se señaló antes, las diferencias raciales existentes en cuanto a CI reflejan, con toda seguridad, diferencias culturales y factores ambientales en grado sustancial. Wilson (1994) ha catalogado las numerosas diferencias en antecedentes culturales entre los estadounidenses de origen europeo y africano. Por ejemplo, en 1992, 64% de padres afroestadounidenses se habían divorciado, separado, habían quedado viudos o nunca se habían casado; 63% de los nacimientos dentro de la comunidad afroestadounidense ocurrieron entre madres solteras y 30% de los nacimientos entre afroestadounidenses fueron de madres adolescentes (*U.S. Bureau of the Census*, 1993). Estas desalentadoras realidades de la vida familiar de muchos afroestadounidenses deben traducirse, inevitablemente, en un desempeño menor en pruebas de inteligencia. Pero, con el propósito de que el lector no concluya que se avala una forma sutil de superioridad anglocéntrica, considérese la conclusión de Lynn (1987) de que la media de CI de los japoneses es de 107, siete puntos por encima del promedio de los individuos de raza banca en EUA. ¿Y qué?

CAMBIOS DE INTELIGENCIA CON LA EDAD

Se volverá ahora la atención hacia otro tema polémico: ¿la inteligencia declina con la edad? Con seguridad, uno de los estereotipos más generalizados acerca del envejecimiento es que con él se pierda capacidad intelectual. Esta idea está tan extendida que pocos legos la cuestionan. Pero debería ponerse en duda.

En general, el estudio empírico de este tema proporciona una conclusión más optimista de la que sugiere el pensamiento común. No obstante, la investigación también revela que los cambios en inteligencia relacionados con la edad son complejos y multifacéticos. La simple pregunta de si "¿la inteligencia declina con la edad?" resulta tener varias respuestas complicadas.

Se puede seguir la evolución de la investigación sobre cambios intelectuales relacionados con la edad de la siguiente manera:

1. Las primeras investigaciones transversales con instrumentos como el WAIS pintaban un cuadro sombrío de disminución lenta de la inteligencia general después de los 15 o 20

años de edad y de una caída precipitada después de los 60 años.

2. Sólo pocos años después, estudios más complejos que utilizaron pruebas secuenciales con instrumentos multidimensionales (como la Prueba de Habilidades Mentales Primarias) sugirieron una trayectoria más optimista para la inteligencia: cambio mínimo en la mayor parte de las capacidades hasta, cuando menos, los 60 años de edad.
3. La investigación paralela que utilizaba la distinción fluida-cristalizada postuló un incremento gradual en la inteligencia cristalizada casi hasta el final de la vida, en yuxtaposición con una disminución rápida en inteligencia fluida.
4. Recientemente, algunos psicólogos han propuesto que la inteligencia del adulto es cualitativamente diferente, similar a una nueva etapa piagetiana que podría llamarse razonamiento posformal. Esta investigación pone en duda la validez ecológica de emplear instrumentos estandarizados con personas de mayor edad. A continuación se analiza con mayor detalle cada una de estas etapas de investigación.

PRIMERAS INVESTIGACIONES TRANSVERSALES

Uno de los primeros estudios amplios acerca de las tendencias con la edad en una prueba de inteligencia de aplicación individual fue el que presentó Wechsler (1944) poco después de la publicación de la Wechsler-Bellevue Forma I. Como ocurre con todas las pruebas Wechsler diseñadas para adultos, las puntuaciones naturales en las subpruebas de la W-B I se transformaban primero en puntuaciones estándar (conocidas como puntuaciones escalares) con media de 10 y desviación estándar de 3. Sin tomar en cuenta la edad de la persona, estas puntuaciones escalares se basaban en un grupo fijo de referencia de 350 participantes entre 20 y 34 años de edad incluidos en la muestra de estandarización. Al consultar la tabla apropiada por edad, la suma de las 11 puntuaciones escalares se utilizaba para encontrar el CI de la persona evaluada.

No obstante, la suma de las puntuaciones escalares en sí misma es un índice directo de la capacidad de un individuo examinado en relación con el grupo de referencia. Wechsler utilizó este índice para proyectar la relación entre edad e inteligencia (figura 7–11). Sus resultados indicaron un rápido crecimiento en la inteligencia general durante la infancia hasta los 15 o 20 años de edad, seguido de una lenta disminución hasta los 65 años. Wechsler se mostró característicamente contundente en la discusión de sus conclusiones:

> Si el hecho de que el crecimiento intelectual cesa aproximadamente a los 15 años de edad ha sido difícil de aceptar, un trago más amargo lo constituye la indicación de que la inteligencia, después de lograr su máximo, comienza a declinar de manera inmediata justo como cualquier otra capacidad fisiológica, en vez de mantenerse en su nivel más alto durante un largo periodo. De hecho, ha resultado tan desagradable, que los psicólogos en general han evitado señalarlo (Wechsler, 1952).

Los estudios normativos con subsecuentes pruebas Wechsler para adultos revelaron justo el mismo patrón, tanto en el WAIS (Wechsler, 1955), el WAIS-R (Wechsler, 1981) y el WAIS-III (Tulsky, Zhu y Ledbetter, 1997). Investigadores posteriores también extendieron el límite superior de edad sometido a prueba a los 70 y 80 años, encontrando una tasa progresiva y acelerada de disminución en el desempeño general en una prueba que parecía confirmar la creencia de Wechsler de que era "normal" que la inteligencia general declinara después de la primera parte de la edad adulta (Eisdorfer y Cohen, 1961).

Wechsler y muchos otros investigadores que emplearon el **diseño transversal** pasaron por alto la influencia de su metodología sobre sus datos. Durante algún tiempo se ha reconocido que los estudios transversales suelen confundir los efectos de la edad con disparidades educativas o con otras diferencias entre grupos por edad (véase Baltes, Reese y Nesselroade, 1977; Kausler, 1991). Por ejemplo, en los estudios

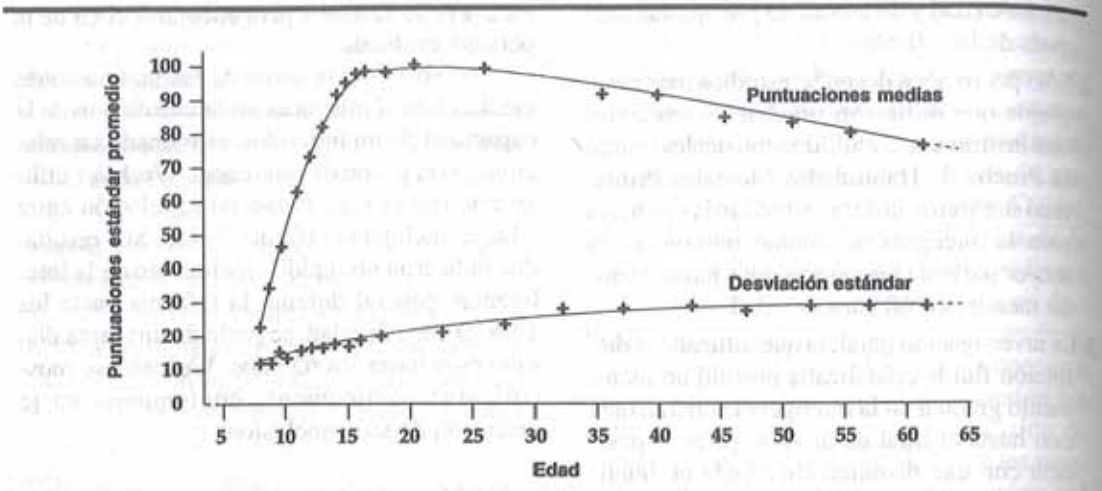


Figura 7–11. Curva de crecimiento y supuesta disminución en la Wechsler-Bellevue Forma I. **Fuente:** Reimpreso con autorización de Wechsler, D. (1944). *Measurement of adult intelligence* (3rd ed.). Baltimore, MD: Williams y Wilkins.

normativos de las pruebas Wechsler, es invariablemente cierto que los individuos más jóvenes de estandarización tienen mayor nivel de educación que los mayores; es muy probable que las bajas calificaciones de estos últimos sean consecuencia, en parte, de estas diferencias educativas, más que significar una disminución inexorable relacionada con la edad.

ESTUDIOS SECUENCIALES DE LA INTELIGENCIA

Para controlar las diferencias entre grupos por edad, muchos investigadores prefieren un **diseño longitudinal**, en el que se somete a prueba a los mismos individuos dos o más veces en periodos de 5 a 10 años y, en casos excepcionales, hasta 40 años después. Debido a que sólo se tiene un grupo de individuos, los diseños longitudinales eliminan las disparidades de los grupos por edad (p. ej., mayor educación en los jóvenes que en las personas mayores) como un factor productor de confusión. No obstante, este enfoque no deja de tener sus defectos:

1. El momento en que ocurre la medición constituye el problema más serio. Los acontecimientos históricos de gran importancia, como una depresión económica, pueden torcer el desarrollo intelectual y psicológico de generaciones completas. Como resultado, los cambios relacionados con la edad que se hayan medido de manera longitudinal pueden reflejar las peculiaridades del momento de la medición, más que cualesquiera efectos universales de la edad.
2. Desgaste selectivo —los individuos menos capaces pueden estar en mayor probabilidad de abandonar el estudio, lo cual infla de manera artificial las calificaciones de la prueba de los individuos que vuelven a examinarse.
3. Efectos de la práctica —los individuos examinados mejoran cuando realizan la misma prueba 2, 3 e incluso 5 veces.
4. Regresión a la media —esto constituye un problema, en especial cuando se seleccio a los participantes debido a sus calificac nes iniciales extremas, como una calificac muy baja de CI (Hayslip y Panek, 1989).

El método más eficaz para el estudio de los cambios en habilidad relacionados con la edad es el **diseño transecuencial**, que combina las metodologías de los estudios transversal y longitudinal (Schaie, 1977):

En pocas palabras, los investigadores comienzan con un estudio transversal. Después, luego de un periodo de años, vuelven a someter a prueba a estos individuos, lo cual proporciona datos longitudinales acerca de varias cohortes —una secuencia longitudinal. Al mismo tiempo, examinan a un nuevo grupo de personas, con lo cual se forma un segundo estudio transversal —y, junto con el primero, una secuencia transversal. El proceso completo puede realizarse una y otra vez (cada 5 o 10 años, por decir) con revaloraciones de los viejos participantes (lo cual aumenta los datos longitudinales) y pruebas iniciales para los nuevos (que aumentan los datos transversales); (Schaie y Willis, 1986).

En 1956, Schaie comenzó el más amplio estudio transecuencial jamás conducido en lo que se conoce como el Estudio longitudinal de Seattle (Schaie, 1958, 1983, 1994). Aplicó la prueba de cinco Habilidades Mentales Primarias (PMA) de Thurstone y otras medidas relacionadas con la inteligencia a una muestra transversal inicial de 500 habitantes adultos de la comunidad. Las subpruebas de la PMA incluyen Significados Verbales, Espacial, Razonamiento, Número y Fluidez de Palabras. En 1963, volvió a someter a prueba a estos individuos y añadió una nueva cohorte transversal. Se recolectaron conjuntos adicionales de datos en 1970, 1977, 1984 y 1991.

El estudio transecuencial de Schaie acerca de las habilidades mentales adultas produjo tres conclusiones:

1. Cada estudio transversal indicó cierto grado de disminución relacionada con la edad en habilidades mentales, comenzando después de los 35 años para algunas capacidades, hasta después de los 50 para otras. En particular, Número y Fluidez Verbal muestran una disminución sólo después de los 50 años, parecen disminuir más pronto, después de los 35 años.
2. Los estudios transversales sucesivos —la secuencia transversal— reveló diferencias intergeneracionales significativas en favor de los más jóvenes. Incluso si se mantiene constante la edad, aquellos nacidos y examinados recientemente tuvieron mejor desempeño que aquéllos nacidos y examinados antes. Por ejemplo, los individuos de 30 años de edad evaluados en 1977 tendieron a obtener mejores calificaciones que las personas de 30 años evaluadas en 1970, quienes tendieron a presentar mejores calificaciones que los individuos examinados en 1963, y éstos, a su vez, tuvieron mejor desempeño que aquéllos examinados en 1956. No obstante, estas diferencias de cohorte en cuanto a inteligencia no fueron uniformes a través de las diferentes habilidades medidas por la PMA. El patrón de habilidades ascendentes fue más aparente para Significados Verbales, Razonamiento y Espacial. Los cambios de cohorte para Número y Fluidez Verbal fueron más complejos y contradictorios.
3. En contraste con las conclusiones moderadamente pesimistas de las comparaciones transversales, las comparaciones **longitudinales** mostraron una tendencia, ya sea a que las calificaciones promedio se eleven de manera imperceptible o que permanezcan constantes hasta los 60 o 70 años de edad. La única excepción en esta tendencia implicó las pruebas que requieren altos niveles de velocidad, como Fluidez de Palabras, en la cual los examinados deben nombrar palabras dentro de una categoría dada con la mayor rapidez posible, y Número, donde las personas deben llevar a cabo cálculos aritméticos de manera rápida y precisa.

Los resultados del estudio de Schaie son incluso más optimistas cuando las conclusiones longitudinales se separan de los promedios de grupo. Como se señaló antes, las conclusiones longitudinales difirieron de una habilidad mental a otra. Sin embargo, al tomar el promedio de las cinco habilidades mentales primarias y utilizar el percentil 25 para los individuos de 25 años de edad como su norma de declinación significativa, Schaie ha mostrado que no más de 25% de aquellas personas estudiadas ha declinado para la edad de 67 años. De los 67 a los 74 años de edad, cerca de un tercio de las personas ha mostrado disminución, mientras que de los 74 a los 81 años apenas un poco más de 40% de las personas ha presentado declinación

(Schaie, 1980, 1984, 1994; Schaie y Willis, 1986). En resumen, la mayoría de personas no muestra una disminución significativa en las habilidades medidas por la Prueba de Habilidades Mentales primarias hasta que se encuentra bastante dentro de su séptimo decenio de vida. Quizás incluso es más impresionante el hecho de que alrededor de 10% de la muestra mejoró de manera significativa cuando se le volvió a examinar en su séptimo y octavo decenio de vida. Con base en su investigación y en otros estudios longitudinales, Schaie llega a esta conclusión:

> Si conserva su salud y participa mentalmente en los problemas y actividades del mundo a su alrededor, es muy probable que experimente poca o ninguna disminución en su desempeño intelectual a lo largo de la vida. Ésa es la promesa de la investigación en el área de la inteligencia adulta (Schaie y Willis, 1986).

Se coincide con esta resolución, pero no sería justo dejar la impresión de que todas las autoridades en el área concuerdan con ella. Horn y Cattell han sido los escépticos que más se han expresado al respecto, afirmando que existe una disminución significativa relacionada con la edad en cuanto a inteligencia fluida, debido a que ésta depende de la integridad neuronal que se supone presenta una declinación a medida que se avanza en edad (Horn y Cattell, 1966; Horn, 1985). Los estudios transversales seguramente apoyan esta perspectiva. Por ejemplo, Wang y Kaufman (1993) graficaron las diferencias de edad en las calificaciones de Vocabulario y Matrices de la *Kaufman Brief Intelligence Test* (Prueba Breve de Inteligencia de Kaufman) y encontraron poco cambio en vocabulario (medida cristalizada), pero una caída pronunciada en Matrices (medida fluida). Estos resultados se mantuvieron incluso cuando se ajustaron las calificaciones en cuanto a nivel educativo (figura 7–12). Por supuesto, los estudios transversales están abiertos a otras interpretaciones y, por tanto, sólo pueden sugerir patrones longitudinales. Los lectores que deseen seguir esta controversia deberían consultar a Kausler (1991) y Lindenberger y Baltes (1994).

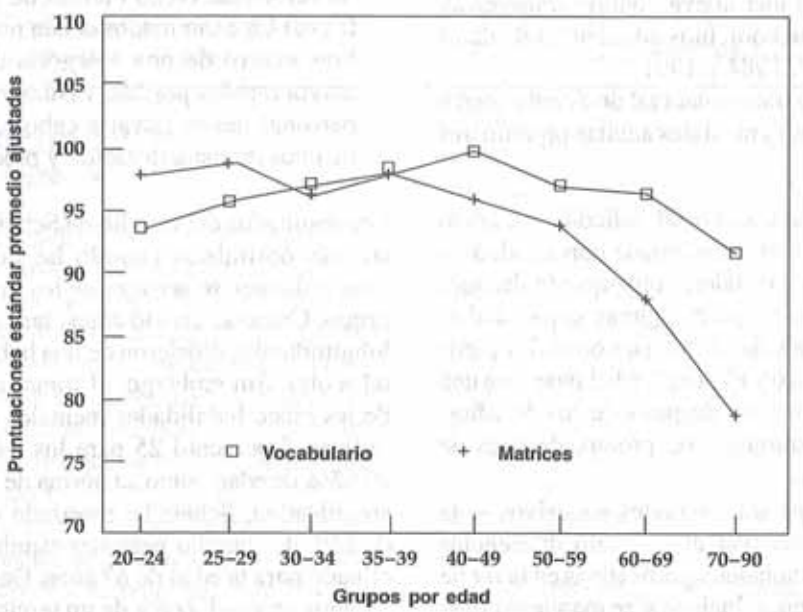


Figura 7–12. Puntuaciones estándar promedio ajustadas en la KBIT en función de la edad.
Fuente: Reproducido con autorización de Wang, J. y Kaufman, A. (1993). Changes in fluid and crystallized intelligence across the 20– to 90– year age range on the K-BIT. *Journal of Psychoeducational Assessment, 11*, 29-37.

CAMBIOS GENERACIONALES EN LAS CALIFICACIONES DE PRUEBAS DE INTELIGENCIA

¿Qué sucede con la inteligencia de una población de una generación a la siguiente? Por ejemplo, ¿cómo se compara la inteligencia de los estadounidenses en el decenio de 1990 con la de sus antepasados a inicios del siglo XX? Podría esperarse que cualesquiera diferencias fueran pequeñas. Después de todo, la reserva genética ha permanecido esencialmente constante durante siglos, quizá milenios. Lo que es más, sólo una pequeña fracción de cualquier generación está expuesta a los extremos de enriquecimiento o privación ambiental que podrían reducir o elevar la inteligencia de manera espectacular. Así, el sentido común dicta que todo cambio generacional en la inteligencia de la población sería mínimo.

Sobre este asunto, el sentido común parece estar en un error. Flynn (1984, 1987, 1994) conjuntó en un cuadro los datos de estandarización de ediciones sucesivas de la Stanford-Binet y de las pruebas Wechsler desde 1932 hasta 1981 y encontró que cada edición estableció una norma más elevada que su antecesora (cuadro 7–6). La ganancia total asciende a una elevación aparente de 13.8 puntos en el CI promedio. Así, una persona que haya obtenido un CI Total de 100 puntos en 1990, habría obtenido en promedio ¡un CI de 114 en 1932! Se ha observado una elevación similar en el desempeño en muchas otras naciones industrializadas utilizando pruebas como Matrices Progresivas de Raven (Flynn, 1987).

En apariencia, los ciudadanos de los países occidentales industrializados han logrado un mejor nivel de educación y de alfabetización en el siglo XX, lo cual ha producido que los niveles del CI promedio se eleven mucho. No obstante, el aumento de CI de esta magnitud plantea un grave problema de explicación causal. Flynn (1994) se muestra escéptico de que cualquier inteligencia real y significativa de una población pueda elevarse con tal rapidez. Concluye que las pruebas actuales no miden inteligencia, sino más bien un correlato con un vínculo causal débil de ésta:

> ... los psicólogos deberían dejar de decir que las pruebas de CI miden inteligencia. Deberían decir que las pruebas de CI miden capacidad para solucionar problemas abstractos (CSPA), un término que transmite de manera precisa nuestra ignorancia. Sabemos que las personas resuelven problemas en las pruebas de CI; sospechamos que éstos están tan separados o alejados de la realidad, que la capacidad para resolverlos puede diferenciarse a través del tiempo de la capacidad de solución de problemas reales llamada inteligencia; aparte de esto sabemos muy poco (Flynn, 1987).

Aunque el precepto radical de Flynn no ha recibido un apoyo general por parte de los expertos en el campo, este autor sí ha sensibilizado a los psicómetras acerca de los peligros de llegar a conclusiones con base en las normas siempre cambiantes de las pruebas de inteligencia. El aumento de los niveles de CI a través del tiempo obliga a que las pruebas vuelvan a estandarizarse con frecuencia; de otro modo, se está calificando a las personas evaluadas con normas obsoletas y se les dan calificaciones de CI infladas (Flynn, 1994). Neisser (1998) ha editado un libro que estudia las posibles explicaciones para la elevación de la curva de puntuaciones de CI.

RESUMEN

1. El sesgo de una prueba se define como la validez diferencial de una interpretación dada a una calificación de prueba para cualquier subgrupo definible y relacionado de personas evaluadas. Como ejemplo del sesgo de una prueba, una prueba verbal de aritmética presentada en inglés podría funcionar bien para niños de habla inglesa,

Cuadro 7–6. Comparación de CI promedio de ediciones sucesivas de las pruebas Stanford-Binet y Wechsler

	Fechas			Edades		Medias		
Combinaciones de prueba	Prueba 1	Prueba 2	Años	Años	N	Prueba 1	Prueba 2	Ganancia
1. SB-L & WISC	1932	1947.5	15.5	5–15	1 563	107.13	101.64	5.49
2. SB-M & WISC	1932	1947.5	15.5	5	46	125.13	107.56	17.57
3. SB-LM & WISC	1932	1947.5	15.5	5–15	460	114.64	109.67	4.97
4. SB-L & WAIS	1932	1953.5	21.5	16–32	271	113.02	105.48	7.54
5. SB-LM & WAIS	1932	1953.5	21.5	16–48	79	109.08	101.75	7.33
6. SB-LM & WPPSI	1932	1964.5	32.5	4–6	416	101.74	92.78	8.96
7. SB-LM & SB-72	1932	1971.5	39.5	2–18	2 351	107.08	97.19	9.89
8. WB-I & WISC	1936.5	1947.5	11	11–14	110	103.51	105.54	-2.03
9. WB-I & WAIS	1936.5	1953.5	17	16–39	152	122.94	118.25	4.69
10. WISC & WAIS	1947.5	1953.5	6	14–17	436	101.76	99.12	2.64
11. WISC & WPPSI	1947.5	1964.5	17	5–6	108	93.56	90.86	2.70
12. WISC & SB-72	1947.5	1971.5	24	6–10	30	96.40	84.42	11.98
13. WISC & WISC-R	1947.5	1972	24.5	6–15	1 042	97.19	88.78	8.41
14. WAIS & WISC-R	1953.5	1972	18.5	16–17	40	102.94	96.29	6.65
15. WAIS & WAIS-R	1953.5	1978	24.5	35–44	72	109.69	101.65	8.04
16. WPPSI & SB-72	1964.5	1971.5	7	4–5	35	93.06	88.65	4.41
17. WPPSI & WISC-R	1964.5	1972	7.5	5–6	140	112.84	108.58	4.26
18. WISC-R & WAIS-R	1972	1978	6	16	80	99.61	98.65	0.96
19. WISC-R & WISC-III	1972	1988	16	6–16	206	108.20	102.90	5.30
20. WAIS-R & WISC-III	1978	1988	10	16	189	105.30	101.40	3.90
21. WPPSI-R & WISC-III	1986	1988	2	6	188	106.50	102.50	4.00
22. WISC-III & WAIS-III	1988	1996	8	16	184	104.60	103.90	.70

Nota: Fechas se refiere a las fechas de estandarización de las pruebas de comparación; **Años** se refiere a los años entre estandarizaciones; **Edades** se refiere al rango de edad de los participantes sometidos a prueba; **Ganancia** se refiere a la ganacia aparente en CI.

Fuente: Adaptado con autorización de Flynn, J. R. (1984). The mean IQ of Americans: Massive Gains 1932 to 1978. *Psychological Bulletin, 95*, 29-51.

mientras que para aquéllos de habla hispana, la misma podría predecir de manera deficiente sus habilidades aritméticas.

2. El sesgo en la validez de contenido se demuestra cuando un reactivo o subescala de una prueba es relativamente más difícil para los miembros de un grupo que para los de otro, después de que se ha mantenido constante el nivel de capacidad general. En términos generales, para las principales pruebas estandarizadas de capacidad y aptitud, la evidencia de sesgo de contenido es escasa o nula.
3. El sesgo en la validez predictiva o relacionada con el criterio se demuestra cuando una prueba no predice igualmente bien un

criterio relacionado en personas de diferentes subpoblaciones. Una prueba imparcial posee regresión homogénea: los resultados de todas las subpoblaciones relacionadas se agrupan por igual alrededor de una sola línea de regresión.

4. El sesgo en la validez de constructo se demuestra cuando una prueba mide rasgos o constructos diferentes en un grupo que en otro. Comparando subpoblaciones relacionadas, una prueba imparcial revelaría un alto grado de semejanza en cuanto a la estructura factorial de la prueba y la clasificación de los reactivos de acuerdo con el nivel de dificultad dentro de la prueba.
5. La justicia de prueba incorpora valores y filosofías sociales acerca del empleo de éstas. Se han descrito tres filosofías: individualismo no cualificado (seleccionar a la mejor persona utilizando todos los medios de predicción), cuotas (selección de acuerdo con proporciones) e individualismo cualificado (seleccionar a la persona sin tomar en cuenta raza, sexo, etc., como medios de predicción). ¿Cuál de las tres filosofías es la correcta? Ésta es una cuestión ética que está más allá de una solución objetiva.
6. Por lo general, la contribución genética a las características humanas se mide en términos de un índice de hereditabilidad que puede variar de 0 a 1.0. La hereditabilidad es un estimado de la cantidad de varianza total de un rasgo dado que se debe a factores genéticos. Ésta es relativa a la población muestreada y no explica las calificaciones individuales. Para el CI, la mayor parte de los estimados de hereditabilidad se encuentran alrededor de .50.
7. La evidencia de una contribución genética a la inteligencia se ha documentado a través de los Estudios Minnesota sobre gemelos, en los que se reúne a gemelos idénticos separados al nacer y se les aplica un examen psicológico extenso. Aunque muchos pares de gemelos se criaron en ambientes disímbolos, sus CI adultos son notablemente similares. Estos datos corroboran estudios anteriores sobre adopción.
8. Jensen (1977) estudió los efectos acumulativos de un ambiente sombrío, de privación, sobre los CI de niños afroestadounidenses de entornos rurales. Encontró que quienes eran mayores tenían calificaciones más bajas que sus hermanos menores, perdiendo aparentemente cerca de un punto de CI por año, en promedio, entre los de 6 y 16 años de edad.
9. Scarr y Weinberg estudiaron los efectos del enriquecimiento ambiental: encontraron que los niños afroestadounidenses adoptados por familias de raza blanca de clase media alta mostraban un CI superior al promedio.
10. El alcoholismo crónico en mujeres embarazadas causa alto riesgo en sus hijos de presentar síndrome alcohólico fetal, que se caracteriza por anormalidades faciales, deficiencias en el crecimiento, problemas motores, hiperactividad y retraso mental de leve a moderado. Con menores niveles de ingestión de alcohol, los hijos pueden mostrar alteraciones en atención y otros problemas sutiles conocidos como efecto alcohólico fetal.
11. Las toxinas ambientales también pueden afectar a la inteligencia. Por ejemplo, los niños que absorben cantidades indebidas de plomo (p. ej., por ingerir pequeños trozos de pintura con pigmentos de plomo) pueden presentar disminuciones a largo plazo en el funcionamiento mental (reducción del CI, problemas con el procesamiento auditivo y del habla y lentificación en el tiempo de reacción).
12. En promedio, los individuos afroestadounidenses obtienen 15 puntos menos que los estadounidenses de raza blanca en pruebas estandarizadas de CI. Cuando se toman en cuenta las variables demográficas como la clase social, la diferencia se reduce de 7 a 10 puntos. En apariencia, la magnitud de la diferencia ha permanecido constante desde mediados hasta finales del siglo XX.
13. Jensen (1969) y otros autores han propuesto que los individuos de raza blanca obtienen mayores calificaciones que los afroestadounidenses en pruebas de CI debido, en

parte, a factores genéticos. Esta hipótesis se basa en la cuestionable suposición de que la evidencia de la hereditabilidad del CI dentro de grupos puede utilizarse para inferirla entre grupos raciales. La investigación sobre mezcla racial y CI no apoya una perspectiva genética.

14. La investigación longitudinal sobre edad e inteligencia proporciona una perspectiva más optimista que la investigación transversal. En los estudios longitudinales, la mayor parte de las capacidades cambian poco hasta que se llega, cuando menos, a los 60 años de edad. Las capacidades fluidas —principalmente la eficiencia mental no verbal y con influencia reducida de la cultura— muestra mayor disminución con la edad que otras capacidades.
15. Flynn conjuntó los datos de estandarización de cada edición de la Stanford-Binet y de las Escalas Wechsler desde 1932 hasta la actualidad. Cada prueba estableció un estándar más alto que su antecesora, con una ganancia total de cerca de 14 puntos de CI, lo que plantea graves problemas de explicación e indica que las normas para las pruebas pueden cambiar con gran rapidez.

TÉRMINOS Y CONCEPTOS CLAVE

Sesgo de una prueba p. 290
Sesgo del contenido p. 293
Sesgo en la validez predictiva p. 294
Sesgo en la validez de constructo p. 297
Equidad de una prueba p. 300
Individualismo no cualificado p. 300
Cuotas p. 301
Individualismo cualificado p. 301
Índice de hereditabilidad p. 302
Teratógenos p. 307
Síndrome alcohólico fetal p. 307
Efecto alcohólico fetal p. 307
Diseño transversal p. 313
Diseño longitudinal p. 314
Diseño transecuencial p. 314–315

■

CAPÍTULO 8

Pruebas grupales de aptitud y de rendimiento

TEMA 8A
Pruebas de aptitud y análisis factorial

Compendio sobre análisis factorial
Baterías de prueba de aptitudes múltiples
Predicción del desempeño en la universidad
Pruebas de selección de posgraduados
Resumen
Términos y conceptos clave

En este capítulo se analiza una variedad de instrumentos que tradicionalmente se agrupan bajo los encabezados de pruebas de aptitud y de rendimiento; incluye instrumentos para este propósito, pero también engloba los temas y aplicaciones en las pruebas de aptitud y rendimiento. En el tema 8A, Pruebas de aptitud y análisis factorial, se describe el empleo del análisis factorial en el desarrollo de las medidas de aptitud; también se revisan los instrumentos típicos, incluyendo las baterías de prueba de aptitudes múltiples y aquellas que se utilizan para predecir el desempeño académico en la universidad. En el tema 8B, Pruebas grupales de rendimiento, se analizan las baterías de prueba de rendimiento educativo familiares para todo alumno del sistema escolar en EUA. Además, el lector encontrará un breve análisis de las pruebas con propósitos especiales para rendimiento, así como una revisión de las cuestiones

sociales perturbadoras que se refieren al falseamiento de resultados dentro del sistema escolar en las pruebas de rendimiento.

Aquí la información se centra en las pruebas de aptitud, especialmente las baterías de aptitudes múltiples que se emplean por lo común para predecir el desempeño futuro en la escuela, empleo y entornos militares. Es común que estas baterías tengan la función de servir como guardianes, pues las tres situaciones anteriores pueden depender de las conclusiones derivadas de las pruebas que se revisan aquí. Las pruebas de aptitud imponen gran respeto y, por tanto, ejercen una inmensa influencia sobre la sociedad moderna; de hecho, su validez es trascendental. El lector aprenderá más sobre la aplicación de estas pruebas en este tema.

Muchas pruebas de aptitud surgieron como derivaciones especializadas de las pruebas de capacidad poco después de que los psicólogos desarrollaron las herramientas estadísticas necesarias para dividir la inteligencia general en sus subcomponentes. Dicho en términos sencillos, la mayor parte de las pruebas de aptitud deben su origen al análisis factorial, una familia de procedimientos que los investigadores utiliza para resumir las relaciones entre variables que se correlacionan de manera sumamente compleja. Debido a que las pruebas de aptitud no pueden florecer sin el análisis factorial, se comenzará esta sección con un compendio de esta útil técnica estadística. Después, el tema continúa con un análisis de las pruebas más importantes de aptitud, que incluyen las baterías de aptitudes múltiples (Test de aptitudes diferenciales, *General Aptitud Test Battery* [Batería de pruebas de aptitud general] y *Armed Services Vocational Assessment Battery* [Batería de evaluación vocacional de las fuerzas armadas]), pruebas que se utilizan para admisión a la universidad (*Scholastic Assessment Tests* [Prueba de evaluación escolar], *American College Test* [Prueba universitaria estadounidense]) y pruebas de admisión al posgrado (*Graduate Record Exam* [Examen de selección para posgrado], *Medical College Admission Test* [Prueba de admisión a la facultad de medicina] y *Law School Admission Test* [Prueba de admisión a la escuela de leyes]).

COMPENDIO SOBRE ANÁLISIS FACTORIAL

En términos amplios, existen dos formas de análisis factorial: confirmatorio y exploratorio. En el primero, el propósito consiste en confirmar que las calificaciones de una prueba y las variables encajan en cierto patrón predicho por una teoría. Por ejemplo, si la teoría subyacente a cierta prueba de inteligencia indica que las subpruebas pertenecen a tres factores (p. ej., factores verbal, de ejecución y de atención), entonces se emprendería un análisis factorial confirmatorio para evaluar la precisión de tal predicción. El análisis factorial confirmatorio es esencial para la validación de muchas pruebas de capacidad.

El propósito central del **análisis factorial** exploratorio es resumir las interrelaciones entre gran cantidad de variables de una manera concisa y precisa como auxiliar para la conceptuación (Gorsuch, 1983). Por ejemplo, el análisis factorial puede ayudar a un investigador a descubrir que una batería de 20 pruebas representa sólo cuatro variables subyacentes, llamadas **factores**. Puede utilizarse el menor conjunto de factores derivados para representar los constructos esenciales que subyacen al grupo completo de variables.

Quizás una analogía simple ayude a aclarar la naturaleza de los factores y su relación con las variables o pruebas de las cuales se derivan. Considérese el decatlón de pista y campo, una mezcla de 10 competencias diversas que incluyen carreras de corta distancia, carreras con obstáculos, salto con garrocha, lanzamiento de bala y carreras de larga distancia, entre otras. Para conceptuar la capacidad del decatleta individual, no se piensa de manera exclusiva en términos de las habilidades del participante para certámenes específicos. En vez de ello, se piensa en términos de atributos más básicos, como velocidad, fortaleza, coordinación y resistencia, cada uno de los cuales se refleja en grado diferente en las competencias individuales. Por ejemplo, el salto con garrocha requiere velocidad y coordinación, mientras las competencias

de carreras de obstáculos demandan coordinación y resistencia. Estos atributos inferidos son análogos a los factores subyacentes al análisis factorial. De la misma manera que los resultados de 10 competencias de un decatlón pueden reducirse a un pequeño número de factores subyacentes (p. ej., velocidad, fortaleza, coordinación y resistencia), los resultados de una batería de 10 o 20 pruebas pueden reflejar la operación de un pequeño número de atributos cognitivos básicos (p. ej., habilidad verbal, visualización, cálculo y atención, por citar una lista hipotética). Este ejemplo ilustra el objetivo del análisis factorial: ayudar a producir una descripción económica de conjuntos amplios y complejos de datos.

Se ilustrarán los conceptos esenciales del análisis factorial con un ejemplo típico relacionado con el número y tipo de factores que pueden describir mejor las capacidades de un estudiante. Holzinger y Swineford (1939) aplicaron 24 pruebas psicológicas relacionadas con capacidad a 145 estudiantes de secundaria de Forest Park Illinois. El análisis factorial que se describe enseguida se basó en los métodos descritos en Kinnear y Gray (1997).

Para el lector debería ser obvio a nivel intuitivo que cualquier batería extensa de pruebas de capacidad reflejará un número menor de capacidades básicas subyacentes (factores). Considérense las 24 pruebas representadas en el cuadro 8–1; seguramente algunas reflejan capacidades subyacentes comunes. Por ejemplo, se esperaría que Oraciones incompletas, Clasificación de palabras y Significados de palabras (variables 7, 8 y 9) midan un factor de capacidad general de lenguaje de algún tipo. De la misma manera, parece probable que otros grupos de pruebas midan capacidades subyacentes comunes. Pero ¿cuántas capacidades o factores?, y ¿cuál es la naturaleza de estas capacidades subyacentes? El análisis factorial es la herramienta ideal para responder a estas preguntas. En este caso se sigue el análisis factorial de los datos de Holzinger y Swineford (1939) de principio a fin.

MATRIZ DE CORRELACIÓN

El punto inicial de todo análisis factorial es la **matriz de correlación**, una tabla completa de intercorrelaciones entre todas las variables.* Las correlaciones entre las 24 variables de capaci-

* En este ejemplo, las variables son pruebas que producen calificaciones más o menos continuas. Pero las variables en un análisis factorial pueden asumir otras formas, en tanto se puedan expresar como calificaciones continuas. Por ejemplo, las siguientes pueden ser variables en un análisis factorial: estatura, peso, ingresos, clase social y resultados en una escala de calificación.

Cuadro 8–1. Las 24 pruebas de capacidad empleadas por Holzinger and Swineford (1939)

1. Percepción visual
2. Cubos
3. Tablero con formas de papel
4. Banderas
5. Información general
6. Comprensión de párrafos
7. Frases incompletas
8. Clasificación de palabras
9. Significado de palabras
10. Sumas de dígitos
11. Claves (velocidad perceptual)
12. Conteo de grupos de puntos
13. Mayúsculas rectas y curvadas
14. Reconocimiento de palabras
15. Reconocimiento de números
16. Reconocimiento de figuras
17. Objeto-número
18. Número-figura
19. Figura-palabra
20. Deducción
21. Acertijos numéricos
22. Razonamiento de problemas
23. Completamiento de series
24. Problemas aritméticos

dad que se analizan aquí pueden encontrarse en el cuadro 8–2. El lector observará que las variables 7, 8 y 9 sí se intercorrelacionan, de hecho, de manera bastante fuerte (correlaciones de .62, .69 y .53), como se sospechaba antes. Este patrón de intercorrelaciones es presunta evidencia de que estas variables miden algo en común; es decir, parece ser que estas pruebas reflejan un factor subyacente común. No obstante, este tipo de análisis factorial intuitivo basado en una inspección visual de la matriz de correlación es muy limitado; simplemente hay demasiadas intercorrelaciones como para que la persona que examina la matriz pueda discernir de manera visual los patrones subyacentes para todas las variables. Aquí es donde puede ser útil el análisis factorial. Aunque no se puede especificar la mecánica del procedimiento, el análisis factorial depende de computadoras de alta velocidad para buscar la matriz de correlación según reglas estadísticas objetivas y determinar el menor número de factores necesarios para explicar el patrón de intercorrelaciones observado. El análisis también produce la matriz factorial, una tabla que muestra el grado en el cual cada prueba tiene cargas en (o se correlaciona con) cada uno de los factores derivados, como se analiza a continuación.

Cuadro 8–2. Matriz de correlación para 24 variables de capacidad

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
2	32																						
3	40	32																					
4	47	23	31																				
5	32	29	25	23																			
6	34	23	27	33	62																		
7	30	16	22	34	66	72																	
8	33	17	38	39	58	53	62																
9	33	20	18	33	72	71	69	53															
10	12	06	08	10	31	20	25	29	17														
11	31	15	09	11	34	35	23	30	28	48													
12	31	15	14	16	22	10	18	27	11	59	43												
13	49	24	32	33	34	31	35	40	28	41	54	51											
14	13	10	18	07	28	29	24	25	26	17	35	13	20										
15	24	13	07	13	23	25	17	18	25	15	24	17	14	37									
16	41	27	26	32	19	29	18	30	24	12	31	12	28	41	33								
17	18	01	18	19	21	27	23	26	27	29	36	28	19	34	35	32							
18	37	26	21	25	26	17	16	25	21	32	35	35	32	21	33	34	45						
19	27	11	31	14	19	25	23	27	27	19	29	11	26	21	19	26	32	36					
20	37	29	30	34	40	44	45	43	45	17	20	25	24	30	27	39	26	30	17				
21	37	31	17	35	32	26	31	36	27	41	40	36	43	18	23	35	17	36	33	41			
22	41	23	25	38	44	39	40	36	48	16	30	19	28	24	25	28	27	32	34	46	37		
23	47	35	38	34	44	43	41	50	50	26	25	35	38	24	26	36	29	27	30	51	45	50	
24	28	21	20	25	42	43	44	39	42	53	41	41	36	30	17	26	33	41	37	37	45	38	43

Nota: Se omiten los decimales.
Fuente: Reproducido con autorización de Holzinger, K., & Harman, H. (1941). *Factor analysis: A synthesis of factorial methods.* Chicago: University of Chicago Press.

MATRIZ FACTORIAL Y CARGAS FACTORIALES

La **matriz factorial** consiste en una tabla de correlaciones llamadas cargas factoriales. Estas últimas que pueden asumir valores desde –1.00 hasta +1.00) indican la importancia de cada variable en cada factor. Por ejemplo, la matriz factorial en el cuadro 8–3 muestra que del análisis se obtuvieron cinco factores (llamados I, II, III, IV y V). Véase que la primera variable, completamiento de series, tiene una fuerte carga positiva de .71 en el factor I, lo cual indica que esta prueba es un índice razonablemente bueno de ese factor. Obsérvese también que esta misma variable tiene una modesta carga negativa de –.11 en el factor II, lo cual indica que, en grado limitado, mide lo opuesto a este factor; es decir, las altas calificaciones en Completamiento de series tienden a significar bajas calificaciones en el factor II y viceversa.

Los factores pueden parecer bastante misteriosos, pero en realidad son muy sencillos a nivel conceptual. Un factor no es nada más que una suma lineal ponderada de las variables; es decir, cada uno es una combinación estadística precisa de las pruebas utilizadas en el análisis. En cierto sentido, un factor es el producto de la "suma" de partes cuidadosamente determinadas de algunas pruebas y quizá de la "resta" de fracciones de otras; lo que hace que sean especiales es el elegante método analítico utili-

Cuadro 8–3. Matriz factorial de ejes principales para 24 variables de capacidad

	Factores				
	I	II	III	IV	V
23. Completamiento de series	.71	–.11	.14	.11	.07
8. Clasificación de palabras	.70	–.24	–.15	–.11	–.13
5. Información general	.70	–.32	–.34	–.04	.08
9. Significado de palabras	.69	–.45	–.29	.08	.00
6. Comprensión de párrafos	.69	–.42	–.26	.08	–.01
7. Frases incompletas	.68	–.42	–.36	–.05	–.05
24. Problemas aritméticos	.67	.20	–.23	–.04	–.11
20. Deducción	.64	–.19	.13	.06	.28
22. Razonamiento de problemas	.64	–.15	.11	.05	–.04
21. Acertijos numéricos	.62	.24	.10	–.21	.16
13. Mayúsculas rectas y curvadas	.62	.28	.02	–.36	–.07
1. Percepción visual	.62	–.01	.42	–.21	–.01
11. Claves (velocidad perceptual)	.57	.44	–.20	.04	.01
18. Número-figura	.55	.39	.20	.15	–.11
16. Reconocimiento de figuras	.53	.08	.40	.31	.19
4. Banderas	.51	–.18	.32	–.23	–.02
17. Objeto-número	.49	.27	–.03	.47	–.24
2. Cubos	.40	–.08	.39	–.23	.34
12. Conteo de grupos de puntos	.48	.55	–.14	–.33	.11
10. Sumas de dígitos	.47	.55	–.45	–.19	.07
3. Tablero de formas de papel	.44	–.19	.48	–.12	–.36
14. Reconocimiento de palabras	.45	.09	–.03	.55	.16
15. Reconocimiento de números	.42	.14	.10	.52	.31
19. Figura-palabra	.47	.14	.13	.20	–.61

zado para obtenerlos. Existen varios métodos que difieren de maneras sutiles, diferencias que salen del propósito de este libro; el lector puede darse una idea de las diferencias al examinar los nombres de los procedimientos: factores de componentes principales, factores de ejes principales, método de mínimos cuadrados no ponderados, método de probabilidad máxima, factorización de imagen y factorización alfa (Tabachnick y Fidell, 1989). La mayor parte de los métodos producen resultados muy similares.

Las cargas factoriales representadas en el cuadro 8–3 no son más que coeficientes de correlación entre variables y factores. Estas correlaciones pueden interpretarse como muestra de la importancia o carga de cada factor en cada variable. Por ejemplo, la variable 9, la prueba de Significado de palabras, tiene una carga muy fuerte (.69) en el factor I, tiene cargas negativas bajas (–.45 y –.29) en los factores II y III, y una carga insignificante (.08 y .001) en los factores IV y V.

REPRESENTACIÓN GEOMÉTRICA DE LAS CARGAS FACTORIALES

Se acostumbra representar los primeros 2 o 3 factores como ejes de referencia en un espacio bidimensional o tridimensional.* Dentro de este esquema pueden representarse las cargas factoriales de cada variable con propósitos de examinación. En el ejemplo se descubrieron cinco factores, demasiados para una visualización simple. No obstante, se puede ilustrar el valor de la representación geométrica al sobresimplificar los datos y representar sólo los primeros dos factores (figura 8–1). En esta gráfica se ha representado cada una de las 24 pruebas contra los dos factores que corresponden a los ejes I y II. El lector observará que las cargas factoriales en el primer factor (I) son uniformemente positivas, mientras las cargas factoriales en el segundo factor (II) constan de una mezcla de cargas positivas y negativas.

MATRIZ FACTORIAL ROTADA

Un punto importante en este contexto es que la posición de los ejes de referencia es arbitraria. No existe nada que impida que el investigador gire los ejes, de modo que produzcan una mejor correspondencia con las cargas factoriales. Por ejemplo, el lector verá en la figura 8–1 que las pruebas 5, 6, 7 y 9 (todas pruebas de lenguaje) se agrupan juntas. Con seguridad se aclararía la interpretación del factor I si se les redirigiera cerca del centro de este agrupamiento (figura 8–2). Esta manipulación también llevaría al factor II junto a las pruebas interpretables 10, 11 y 12 (todas pruebas numéricas).

Aunque la rotación puede conducirse de manera manual a través de una inspección visual, es más común que los investigadores dependan de uno o más criterios estadísticos para producir la matriz factorial rotada final. Por lo común, se aplican los criterios de Thurstone (1947) de variedad positiva y estructura simple. En una **rotación hacia variedad positiva**, el programa de cómputo busca eliminar la mayor cantidad posible de cargas factoriales negativas. Éstas tienen poco sentido en la evaluación de la capacidad, dado que implican que las altas calificaciones en un factor se correlacionan con un bajo desempeño en prueba. En una **rotación hacia estructura simple**, el programa de cómputo busca simplificar las cargas factoriales, de modo que cada prueba tenga cargas significativas en el menor número de factores posible. La meta de ambos criterios es producir una matriz factorial rotada tan sencilla y sin ambigüedades como sea posible.

En el cuadro 8–4 se muestra la matriz factorial rotada para este problema. El método particular de rotación que se utiliza aquí se denomina rotación varimax; no debe utilizarse si la expectativa teórica sugiere que puede haber un factor general. ¿Se debería esperar un factor general en el análisis de las pruebas de capacidad? La respuesta es que esto es en mucho una cuestión de fe al igual que de ciencia. Un investi-

* Técnicamente es posible representar todos los factores como ejes de referencia en un espacio de n dimensiones, donde n es el número de factores. No obstante, cuando se trabaja con más de 2 o 3 ejes de referencia, ya no es posible hacer una representación visual.

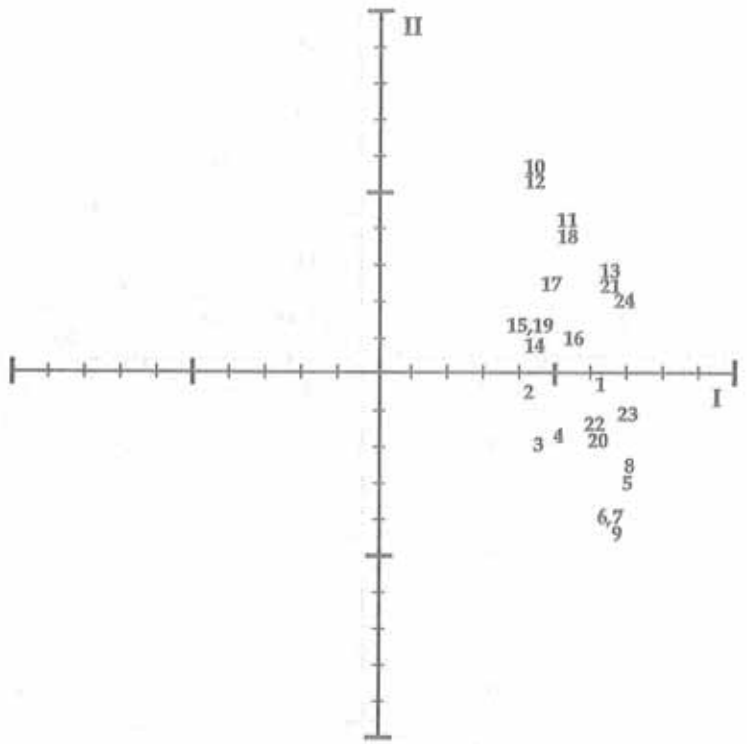


Figura 8–1. Representación geométrica de los primeros dos factores de 24 pruebas de capacidad.

gador puede concluir que es probable la existencia de un solo factor y, por tanto, seguir un tipo diferente de rotación. Un segundo investigador puede sentirse cómodo con una perspectiva similar a la de Thurstone y buscar factores múltiples de capacidad utilizando una rotación varimax. Después se estudiará esta cuestión en detalle, pero es necesario señalar que un investigador se enfrenta a muchos momentos de decisión al realizar un análisis factorial. No es de sorprender, entonces, que diferentes investigadores puedan llegar a diversas conclusiones a partir del análisis factorial, incluso cuando analizan el mismo conjunto de datos.

INTERPRETACIÓN DE FACTORES

El cuadro 8–4 indica que cuatro factores subyacen a las intercorrelaciones de las 24 pruebas de capacidad; pero, ¿cómo se le llamará a estos factores? El lector puede encontrar inquietante la respuesta a esta pregunta, dado que en este punto se dejan de lado los datos estadísticos fríos y objetivos para entrar en el área del criterio, el discernimiento y las suposiciones. A fin de interpretar o nombrar a un factor, el investigador debe realizar un juicio razonado acerca de los procesos y capacidades comunes que comparten las pruebas con fuertes cargas en ese factor. Por ejemplo, en el cuadro 8–4 se muestra que el factor I tiene que ver con la capacidad verbal, dado que las variables con **altas** cargas destacan la habilidad verbal (p. ej., Frases incompletas tiene una carga de .86, Significado de palabras tiene una carga de .84 y Significados de párrafos tiene una carga de .81). Las variables con cargas bajas también ayudan a afinar el significado del factor I. Por ejemplo, el factor I no se relaciona con habilidad numérica (Acertijos numéricos tiene una carga de .18) o habilidad espacial (Tablero con formas de

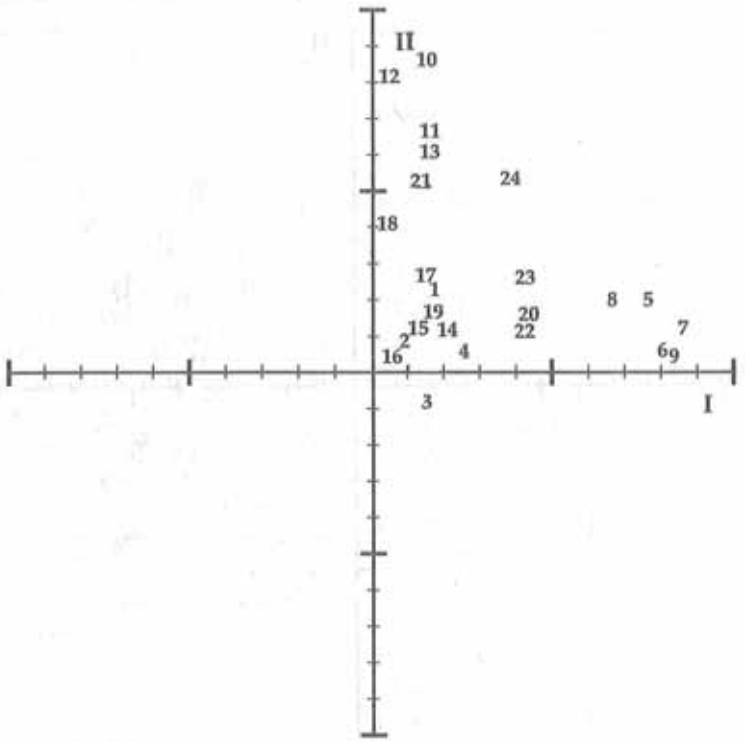


Figura 8–2. Representación geométrica de los primeros dos factores rotados de las 24 pruebas de capacidad.

papel tiene una carga de .16). Si se utiliza una forma similar de inferencia, parece ser que el factor II se relaciona con la capacidad numérica (Sumas tiene una carga de .85, Conteo de grupos de puntos tiene una carga de .89). El factor III es menos seguro, pero parece relacionarse con capacidad visual y el factor IV parece ser una medida del reconocimiento. Se necesitaría analizar la única prueba en el factor V (Figura-Palabra) para conjeturar el significado de este factor.

Estos resultados ilustran uno de los principales usos del análisis factorial, es decir, la identificación de un pequeño número de pruebas indicadoras de una batería grande de pruebas. En vez de utilizar una batería engorrosa de 24 pruebas, un investigador podría obtener casi la misma información al seleccionar con cuidado varias pruebas con fuertes cargas factoriales en los cinco factores. Por ejemplo, el primer factor está bien representado en la prueba 7: Frases incompletas (.86) y la prueba 9: Significados de palabras (.84); el segundo se refleja en la prueba 10: Sumas (.85), mientras el tercero se ilustra mejor a través de la prueba 1: Percepción visual (.69). El cuarto factor es representado por la prueba 15, Reconocimiento de Números (.74) y Reconocimiento de Palabras (.69). Por supuesto, el último factor sólo tiene cargas adecuadas en la prueba 19, Figura-Palabra (.77).

TEMAS SOBRE EL ANÁLISIS FACTORIAL

Por desgracia, es frecuente que el análisis factorial se preste a malos entendidos y a menudo se le utiliza de manera inadecuada. Algunos investigadores parecen emplearlo como una especie de varita mágica, con la esperanza de encontrar el oro oculto bajo toneladas de lodo. Pero la técnica no tiene nada de mágico. Ninguna cantidad de análisis estadístico puede rescatar

Cuadro 8–4. Matriz factorial con rotación varimax para 24 variables de capacidad

	Factores				
	I	II	III	IV	V
7. Frases incompletas	**.86**	.15	.13	.03	.07
9. Significado de palabras	**.84**	.06	.15	.18	.08
6. Comprensión de párrafos	**.81**	.07	.16	.18	.10
5. Información general	**.79**	.22	.16	.12	−.02
8. Clasificación de palabras	**.65**	.22	.28	.03	.21
22. Razonamiento de problemas	**.43**	.12	.38	.23	.22
10. Sumas de dígitos	.18	**.85**	−.10	.09	−.01
12. Conteo de grupos de puntos	.02	**.80**	.20	.03	.00
11. Claves (velocidad perceptual)	.18	**.64**	.05	.30	.17
13. Mayúsculas rectas y curvadas	.19	**.60**	.40	−.05	.18
24. Problemas aritméticos	.41	**.54**	.12	.16	.24
21. Acertijos numéricos	.18	**.52**	.45	.16	.02
18. Número-figura	.00	**.40**	.28	.38	.36
1. Percepción visual	.17	.21	**.69**	.10	.20
2. Cubos	.09	.09	**.65**	.12	−.18
4. Banderas	.26	.07	**.60**	−.01	.15
3. Tablero con formas de papel	.16	−.09	**.57**	−.05	.49
23. Completamiento de series	.42	.24	**.52**	.18	.11
20. Deducción	.43	.11	**.47**	.35	−.07
15. Reconocimiento de números	.11	.09	.12	**.74**	−.02
14. Reconocimiento de palabras	.23	.10	.00	**.69**	.10
16. Reconocimiento de figuras	.07	.07	.46	**.59**	.14
17. Objeto-número	.15	.25	−.06	**.52**	.49
19. Figura-palabra	.16	.16	.11	.14	**.77**

Nota: las cifras en negritas significan subpruebas con una fuerte carga en cada factor.

los datos basados en medidas triviales, irrelevantes o fortuitas. Si no existe oro que encontrar, entonces no se encontrará nada: el análisis factorial no es alquimia; producirá resultados significativos sólo cuando la investigación haya sido significativa desde un principio.

Un aspecto importante es que sólo puede surgir un tipo particular de factor a partir del análisis factorial si, para empezar, las pruebas y medidas contienen ese factor. Por ejemplo, no es posible que surja un factor de memoria a corto plazo a partir de una batería de prueba de capacidad, si ninguna de la prueba requiere de ese tipo de memoria. En general, la calidad de la información obtenida depende de la calidad de la información que se ingresa. Se puede volver a destacar este punto con las siglas BEBS "basura que entra, basura que sale".

La magnitud de la muestra es importante para un análisis factorial estable. Comrey (1973) ofrece la siguiente guía aproximada:

Magnitud de la muestra	Clasificación
50	muy deficiente
100	deficiente
200	adecuada
300	buena
500	muy buena
1 000	excelente

En general, es reconfortante tener cuando menos cinco personas por cada prueba o variable (Tabachnick y Fidell, 1989).

Por último, no se puede enfatizar el grado en que el análisis factorial se guía por las elecciones subjetivas y los prejuicios teóricos. Una cuestión importante al respecto es la elección entre los ejes ortogonales y los ejes oblicuos. Con los **ejes ortogonales**, los factores se encuentran en ángulos rectos entre sí, lo cual significa que no están correlacionados (las figuras 8–1 y 8–2 presentan este tipo de ejes). En muchos casos, los agrupamientos de cargas factoriales están situados de tal manera, que los **ejes oblicuos** proporcionan una mejor correspondencia. Con éstos, los factores se correlacionan entre sí. Algunos investigadores sostienen que siempre deberían utilizarse estos ejes, mientras otros asumen un enfoque más experimental. Tabachnick y Fidell (1989) recomiendan una estrategia exploratoria basada en análisis factoriales repetidos. Su enfoque es descaradamente oportunista:

> Durante algunas secuencias subsecuentes, los investigadores experimentan con diferentes números de factores, diferentes técnicas de extracción y rotaciones tanto ortogonales como oblicuas. Cierto número de factores con alguna combinación de extracción y rotación produce la solución con la mayor utilidad, consistencia y significado científicos; ésta es la solución que se interpreta.

Con las rotaciones oblicuas también es posible realizar un análisis factorial de los factores en sí. Dicho procedimiento puede producir uno o más factores de segundo orden, que pueden apoyar la organización jerárquica de los rasgos y ofrecer un acercamiento entre los teóricos que postulan un solo factor general (p. ej., Spearman) y aquellos que promueven varios factores grupales (p. ej., Thurstone). Quizás ambos bandos estén en lo correcto, con los factores grupales colocados por debajo de un factor general de segundo orden.

BATERÍAS DE PRUEBA DE APTITUDES MÚLTIPLES

Como se señaló antes, las pruebas de aptitud no florecieron hasta que se tuvieron disponibles las herramientas estadísticas —procedimientos de análisis factorial— que constituían un prerrequisito. Una de las principales aplicaciones del análisis factorial fue el desarrollo de las baterías de prueba de aptitudes múltiples, en las cuales se examina a la persona en varias áreas independientes y homogéneas de aptitud. El desarrollo de las subpruebas se inspira en las conclusiones del análisis factorial. Por ejemplo, Thurstone desarrolló una de las primeras baterías de prueba de aptitudes múltiples, la Prueba de Habilidades Mentales Primarias, un conjunto de siete pruebas elegidas con base en el análisis factorial (Thurstone, 1938).

Recientemente, varias baterías de aptitudes múltiples han obtenido popularidad para la orientación educativa y vocacional, asignación vocacional y clasificación en las fuerzas armadas (Gregory, 1994a). Cada año, a cientos de miles de personas se les aplica una de estas importantes baterías: el Test de Aptitudes Diferenciales (DAT), la *General Aptitude Test Battery*, GATB (Batería de prueba de aptitud general) y la *Armed Services Vocational Aptitude Battery*, ASVAB (Batería de aptitud vocacional de las fuerzas armadas). Estas baterías utilizaron el análisis factorial, ya fuera de manera directa para establecer las subpruebas útiles o se guiaron en su construcción de los resultados acumulados en otras investigaciones analítico-factoriales. Las características sobresalientes de cada batería se resumen brevemente en las siguientes secciones.

TEST DE APTITUDES DIFERENCIALES (DAT)

El DAT se publicó inicialmente en 1947 para proporcionar una base para la orientación educativa y vocacional de los alumnos de bachillerato. Después, los examinadores encontraron que esta prueba es útil en la orientación vocacional de jóvenes adultos que se encuentran fuera de la escuela y en la selección de empleados. La prueba ha tenido revisiones periódicas y es una de las baterías de prueba de aptitudes múltiples más popular de todos los tiempos (Bennett, Seashore y Wesman, 1982, 1984).

El DAT incluye ocho pruebas independientes:

1. Razonamiento Verbal (RV).
2. Razonamiento Numérico (RN).
3. Razonamiento Abstracto (RA).
4. Velocidad y Precisión Perceptual (VPP).
5. Razonamiento Mecánico (RM).
6. Relaciones Espaciales (RE).
7. Ortografía (O).
8. Uso del lenguaje (UL).

En la figura 8–3 se muestra un reactivo característico de cada prueba.

Los autores eligieron las áreas para las ocho pruebas con base en datos experimentales y experienciales, más que en un análisis factorial formal realizado por ellos mismos. En la construcción del DAT, los autores se guiaron por varios criterios explícitos:

- Cada prueba debe ser independiente: existen situaciones en las que se requiere o se desea sólo una parte de la batería.
- Las pruebas deben medir poder: para la mayor parte de los propósitos vocacionales a los que contribuyen los resultados de una prueba, la evaluación del poder —resolver problemas difíciles en el tiempo adecuado— es de interés primordial.
- La batería de prueba debe producir un perfil: las ocho calificaciones independientes pueden convertirse en rangos percentiles y representarse en una gráfica común de perfil.
- Las normas deben ser adecuadas: en la quinta edición, las normas se obtuvieron de 100 000 estudiantes para la estandarización de otoño y 70 000 para la de primavera.
- Los materiales de prueba deben ser prácticos: con límites de tiempo de 6 a 30 minutos por prueba, el DAT completo puede aplicarse en una sesión escolar matutina o vespertina.
- Las pruebas deberían ser fáciles de aplicar: cada una contiene excelentes ejemplos de "calentamiento" y las pueden aplicar personas con un mínimo de entrenamiento especial.
- Deben estar disponibles formas alternas: con propósitos de repetición de prueba, la disponibilidad de formas alternas (en la actualidad formas C y D) reducirá cualesquier efectos de la práctica.

La confiabilidad del DAT es, por lo general, bastante alta, con coeficientes por mitades que en gran medida se localizan en el rango de .90 y confiabilidades de formas alternas que se encuentran en el rango de .73 a .90, con mediana de .83. Razonamiento mecánico es una excepción, con confiabilidades tan bajas como .70 en el caso de las niñas. Las pruebas muestran un patrón mixto de intercorrelaciones, que los autores interpretan de manera optimista como una confirmación de la independencia de las ocho pruebas. En realidad, muchas de las

RAZONAMIENTO VERBAL

Elija el par correcto de palabras para llenar los espacios en blanco.

La _____ es a la vista lo que el tímpano es al _____

A. visión — sonido
B. iris — audición
C. retina — oído
D. vista — cóclea
E. pestaña — lóbulo

CAPACIDAD NUMÉRICA

Elija la respuesta correcta.

4(–5)(–3)=

A.–60 B. 27 C.–27 D. 60 E. Ninguna de las anteriores

Figura 8–3. Test de Aptitudes diferenciales y Reactivos característicos.

RAZONAMIENTO ABSTRACTO

Las cuatro figuras en la fila a la izquierda constituyen una serie. Encuentra la única opción a la derecha que continuaría la serie.

< <> <<>> <<>>>> <> <<<>> <<<>>>> <<<<>>>>

A B C D

VELOCIDAD Y PRECISIÓN EN TRABAJOS DE OFICINA

En cada reactivo de la prueba, una de las combinaciones se encuentra subrayada. Marque la misma combinación en la hoja de respuesta.

1.	AB	Ab	AA	BA	Bb	2.	5m	5M	M5	Mm	m5
	Ab	Bb	AA	BA	AB		M5	m5	Mm	5m	5M
1.	O	O	O	O	O	2.	O	O	O	O	O

RAZONAMIENTO MECÁNICO

¿Cuál palanca requerirá mayor fuerza para levantar un objeto del mismo peso? De ser igual, marque C.



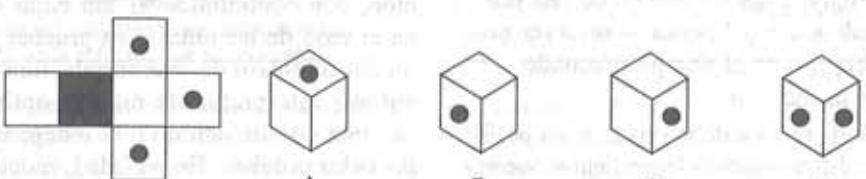
A C (igual)



B

RELACIONES ESPACIALES

¿Cuál de las figuras de la derecha se puede hacer al doblar el patrón a la izquierda? El patrón siempre representa la parte externa de la figura.



A B C D

ORTOGRAFÍA

Marque si cada palabra está escrita de manera correcta o incorrecta.

1. irelevante C I
2. parsimonioso C I
3. Exselente C I

USO DEL LENGUAJE

Decida qué parte de la oración contiene un error y marque la letra correspondiente en la hoja de respuesta. Marque N (ninguna) si no existe error.

A pesar de la crítica del público/ el investigador estudió/

A B

los afectos de la radiación/ sobre el crecimiento de las plantas.

C D

Figura 8–3. Test de Aptitudes diferenciales y Reactivos característicos (continuación).

correlaciones son bastante altas y parece probable que las ocho pruebas reflejen un menor número de factores de capacidad. Con toda seguridad, las pruebas de Razonamiento Verbal y Capacidad numérica miden un solo factor sólido, con correlaciones alrededor de .70 en diversas muestras.

El manual presenta amplios datos que demuestran que las pruebas del DAT, en especial la combinación RV + RN, constituyen buenos medios de predicción de las calificaciones en educación media superior y universidad (correlaciones en los rangos de .60 y .70). Por tal razón, la combinación de RV + RN se considera con frecuencia como un índice de la aptitud escolar. La evidencia de la validez **diferencial** de las otras pruebas es bastante limitada. Bennett, Seashore y Wesman (1974) sí presentan los resultados de varios estudios de seguimiento en los que se correlaciona el ingreso a la universidad/éxito con los perfiles del DAT, pero sus métodos de investigación son más impresionistas que cuantitativos; el observador independiente tendrá dificultades para utilizar sus resultados. Schmitt (1995) señala que uno de los principales problemas con la batería es la

> ... falta de validez discriminativa entre las ocho subpruebas. Con excepción de la prueba de Velocidad y Precisión Perceptual, todas las subescalas tienen fuertes intercorrelaciones (.50 a .75). Si sólo se quiere obtener un índice general de la capacidad académica de la persona, esto es adecuado; si las calificaciones en las subpruebas han de utilizarse en algún sentido diagnóstico, este nivel de intercorrelación hace que las afirmaciones sobre las fortalezas y debilidades relativas de los estudiantes sean altamente cuestionables.

Aun así, el DAT revisado es mejor que las ediciones anteriores. Una mejoría importante es la eliminación del aparente sesgo de género en las pruebas de Uso del Lenguaje y Razonamiento Mecánico —una fuente de críticas en ediciones anteriores. Los editores de la Quinta Edición llevaron a cabo un cuidadoso estudio sobre sesgo de género, raza, étnico y regional, a fin de eliminar los reactivos que podrían ser estereotípicos u ofensivos para subgrupos identificables. Lo que es más, realizaron análisis estadísticos complejos para eliminar reactivos de dificultad desigual entre subgrupos raciales equiparados en términos de desempeño general.

BATERÍA DE PRUEBA DE APTITUD GENERAL (GATB)

A finales del decenio de 1930, el Departamento del Trabajo de EUA desarrolló pruebas de aptitud para predecir el desempeño laboral en 100 ocupaciones específicas. En el decenio de 1940, el departamento contrató a un panel de expertos en medición y psicología industrial-organizacional para crear una batería de prueba de aptitudes múltiples que evaluara las 100 ocupaciones antes estudiadas y muchas más. El resultado de este esfuerzo titánico fue la *General Aptitude Test Battery*, GATB (Batería de Prueba de Aptitud General), ampliamente conocida como la principal batería de prueba para la predicción del desempeño en el trabajo (Hunter, 1994).

La GATB se obtuvo de un análisis factorial de 59 pruebas aplicadas a miles de aprendices varones en cursos vocacionales (*United States Employement Service*, 1970). Las normas interpretativas se han revisado y actualizado de manera periódica, de modo que la GATB es un instrumento totalmente moderno, aunque su contenido haya cambiado muy poco. Una limitación es que la batería está disponible principalmente para oficinas gubernamentales de empleo, aunque organizaciones no lucrativas, incluyendo bachilleratos y ciertas universidades, pueden llegar a arreglos especiales para utilizarla.

La GATB se integra de ocho pruebas de lápiz y papel y cuatro medidas con aparatos. La batería completa puede aplicarse en aproximadamente dos horas y media y es apropiada para estudiantes de bachillerato y adultos. Las 12 pruebas producen un total de nueve calificaciones factoriales:

- **Capacidad general de aprendizaje** (inteligencia) (G). Esta calificación es una combi-

nación de Vocabulario, Razonamiento aritmético y Espacio tridimensional.

- **Aptitud verbal** (V). Obtenida de la prueba de Vocabulario que requiere que la persona indique las dos palabras que en un conjunto son sinónimos o antónimos.
- **Aptitud numérica** (N). Esta calificación es una combinación de las pruebas de Cálculo y Razonamiento Numérico.
- **Aptitud espacial** (E). Consta de la prueba de Espacio Tridimensional, una medida de la capacidad para percibir representaciones bidimensionales de objetos tridimensionales y de visualizar el movimiento en tres dimensiones.
- **Percepción de la forma** (P). Esta calificación es una combinación de Equiparación de Formas y Equiparación de Herramientas, dos pruebas en las que la persona debe equiparar dibujos idénticos.
- **Percepción en trabajos de oficina** (Q). Una prueba de corrección llamada Comparación de nombres en la que el individuo tiene que equiparar nombres bajo presión de tiempo.
- **Coordinación motora** (K). Mide la capacidad para hacer con rapidez marcas específicas con un lápiz en la prueba de Marcado.
- **Destreza digital** (D). Una combinación de las calificaciones de las pruebas de Armado y Desarmado, dos medidas de destreza con remaches y arandelas.
- **Destreza manual** (M). Una combinación de las calificaciones de las pruebas de Colocar y Girar, dos pruebas que requieren que la persona transfiera e invierta clavijas en un tablero.

Las nueve calificaciones factoriales en la GATB se expresan como puntuaciones estándar con una media de 100 y DE de 20. Estas puntuaciones estándar se sustentan en la muestra normativa original de 4 000 trabajadores que se obtuvo en el decenio de 1940. Los coeficientes de confiabilidad de formas alternas van desde el rango de .80 hasta el de .90. El manual de la GATB resume varios estudios de validez acerca de la prueba, principalmente en términos de su correlación con medidas pertinentes de criterio. Hunter (1994) señala que las calificaciones de la GATB predicen el éxito del entrenamiento para todos los niveles de complejidad laboral. El coeficiente promedio de validez es un fenomenal .62.

Las calificaciones absolutas son de menor interés que su comparación con los Patrones de Aptitud Laboral (PAL) actuales para docenas de trabajos. Con base en los resultados de grandes muestras de solicitantes y empleados en diversas ocupaciones, los orientadores y empleadores pueden tener acceso en la actualidad a un cúmulo de información acerca de patrones de calificación que son necesarios para el éxito en una variedad de puestos. Así, una manera de utilizar la GATB consiste en comparar las calificaciones de una persona examinada con los PAL que se consideran necesarios para la destreza en diversas ocupaciones.

Hunter (1994) recomienda una estrategia alternativa basada en aptitudes combinadas (figura 8–4). Las nueve puntuaciones factoriales específicas se combinan muy bien en tres factores generales: cognitivo, perceptual y psicomotor. Hunter señala que los diferentes puestos requieren diversas capacidades de estas aptitudes. Por ejemplo, un obrero en una línea de ensamblado en una planta automotriz podría necesitar altas calificaciones en las combinaciones psicomotora y perceptual, mientras la puntuación cognitiva sería menos importante para su trabajo. La investigación de Hunter demuestra que los factores generales dominan sobre los específicos en la predicción del desempeño laboral. Davison, Gasser y Ding (1996) estudian enfoques adicionales para el análisis e interpretación del perfil de la GATB.

Van de Vijver y Harsveld (1994) investigaron la equivalencia de su versión por computadora de la GATB con la versión tradicional de lápiz y papel. Por supuesto, sólo se compararon las subpruebas cognitivas y perceptuales —las pruebas de habilidad motriz no pueden trasladarse a computadora. Estos autores encontraron que las dos versiones no eran equivalentes. En particular, las subpruebas por computadora producían respuestas más rápidas e imprecisas que

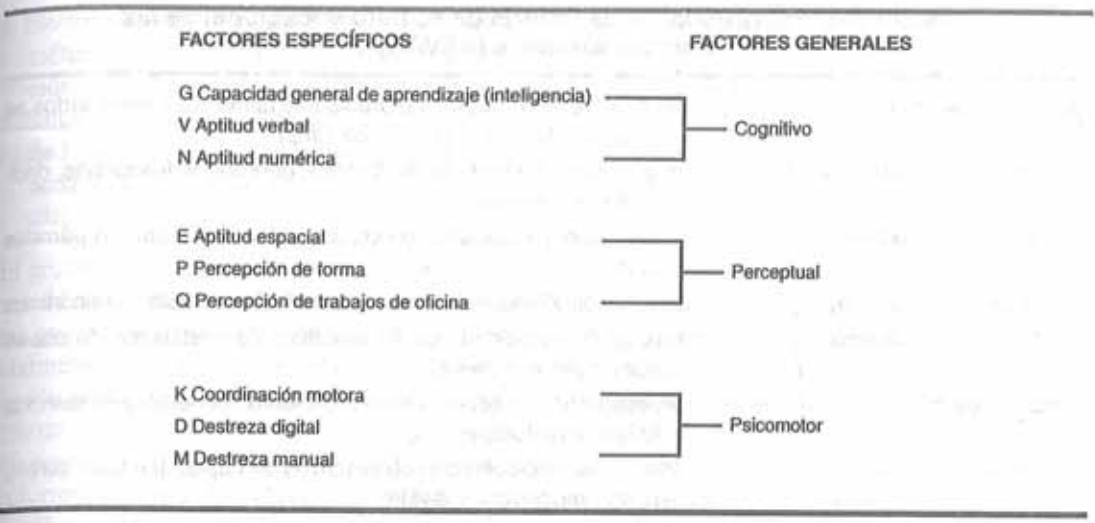


Figura 8–4. Factores específicos y generales en la GATB.

las subpruebas convencionales. Su investigación demuestra una vez más que no puede suponerse la equivalencia entre las versiones tradicional y por computadora de una prueba. Ésta es una pregunta empírica que sólo puede responderse a través de investigación cuidadosa. Nijenhuis y van der Flier (1997) estudian una versión holandesa de la GATB y su aplicación para el estudio de diferencias cognitivas entre inmigrantes y miembros del grupo mayoritario en los Países Bajos.

BATERÍA DE APTITUD VOCACIONAL DE LAS FUERZAS ARMADAS

La *Armed Services Vocational Aptitude Battery*, ASVAB (Batería de Aptitud Vocacional de las Fuerzas Armadas) es probablemente la prueba de lápiz y papel de uso más amplio que se tenga en existencia. Las Fuerzas Armadas la utilizan para seleccionar a reclutas potenciales y asignar personal a diferentes trabajos y programas de entrenamiento. Más de dos millones de personas responden la ASVAB cada año. La versión actual consiste en 10 subpruebas, cuatro de las cuales producen la *Armed Forces Qualification Test*, AFQT (Prueba de Capacitación para las Fuerzas Armadas), el examen común de capacidad para todos los servicios dentro del ejército (cuadro 8–5). Ocho subpruebas son de poder con límites de tiempo adecuados para la mayoría de las personas, mientras dos (Operaciones Numéricas y Velocidad de Codificación) son con límite de tiempo que le dan gran importancia al desempeño rápido. Los coeficientes de confiabilidad de formas alternas para las calificaciones de la ASVAB se encuentran aproximadamente entre .85 y .95, y los coeficientes *test-retest* van desde mitad del rango de .70 hasta mitad del rango de .80 (Larson, 1994). La única excepción es Comprensión de Párrafos, con una confiabilidad de sólo .50. La prueba tiene buenas normas basadas en una muestra representativa de 12 000 personas entre los 16 y 23 años de edad. El manual de la ASVAB presenta un coeficiente mediano de validez de .60 con medidas del desempeño en el entrenamiento.

Las decisiones acerca de las personas examinadas con la ASVAB se basan por lo común en puntuaciones mixtas, no en calificaciones de subprueba. Por ejemplo, la puntuación mixta de trabajos de oficina se obtiene de la combinación de Conocimiento de Palabras, Comprensión de Párrafos, Operaciones Numéricas y Velocidad de Codificación. Los individuos que tienen altas calificaciones en esta combinación podrían asignarse a puestos secretariales. Dado que las puntuaciones mixtas se obtienen a nivel

Cuadro 8–5. Subpruebas de la Batería de Aptitud Vocacional de las Fuerzas Armadas (ASVAB)

Razonamiento aritmético*	prueba con 30 reactivos de problemas aritméticos expresados en palabras, basados en cálculo simple
Conocimientos matemáticos*	prueba con 25 reactivos de álgebra, geometría, fracciones, decimales y exponentes
Comprensión de párrafos*	prueba con 15 reactivos de comprensión de lectura en párrafos cortos
Conocimiento de palabras*	prueba con 35 reactivos de conocimiento de vocabulario y sinónimos
Velocidad de codificación	prueba de velocidad con 84 reactivos de sustitución de claves numéricas por verbales
Ciencia general	prueba con 25 reactivos de conocimientos generales en ciencias físicas y biológicas
Operaciones numéricas	prueba de velocidad con 50 reactivos de capacidad para sumar, restar, multiplicar y dividir
Información sobre electrónica	prueba con 20 reactivos sobre principios de electrónica, radio y electricidad
Comprensión mecánica	prueba con 25 reactivos sobre principios mecánicos y físicos
Información automotriz y mecánica	prueba con 25 reactivos de conocimientos básicos sobre automóviles, prácticas en un taller y utilización de herramientas

* *Armed Forces Qualifying Test* AFQT (Prueba de Aptitudes para las Fuerzas Armadas).

empírico, en cualquier momento pueden desarrollarse nuevas calificaciones para la toma de decisiones sobre asignación a un trabajo. Las puntuaciones mixtas se actualizan y revisan continuamente. Por ejemplo, Ree y Carretta (1994) defienden la presencia de tres puntuaciones mixtas obtenidas de un análisis factorial de más de 11 000 participantes en la evaluación con la ASVAB. Éstas y las subpruebas que las constituyen son las siguientes:

1. Velocidad: Operaciones Numéricas y Velocidad de Codificación.
2. Verbal/Matemáticas: Razonamiento Aritmético, Conocimiento de Palabras, Comprensión de Párrafos y Conocimiento Matemático.
3. Conocimiento técnico: Ciencia General, Información Automotriz y Mecánica, Comprensión Mecánica e Información sobre Electrónica.

El lector observará que el segundo factor es igual al de la AFQT, mencionado antes.

En algún momento, las Fuerzas Armadas dependieron en gran medida de las siete puntuaciones mixtas que se listan a continuación (Murphy, 1984). Las primeras tres constituyen puntuaciones mixtas académicas, mientras las restantes son puntuaciones mixtas laborales. El lector verá que las subpruebas individuales pueden aparecer en más de una combinación:

1. Capacidad académica: Conocimiento de Palabras, Comprensión de Párrafos y Razonamiento Aritmético.
2. Verbal: Conocimiento de Palabras, Comprensión de Párrafos y Ciencia General.
3. Matemáticas: Conocimiento Matemático y Razonamiento Aritmético.
4. Mecánica y oficios: Razonamiento Aritmético, Comprensión Mecánica, Información Mecánico-Automotriz e Información sobre Electrónica.
5. Negocios y oficina: Conocimiento de Palabras, Comprensión de Párrafos, Conocimiento Matemático y Velocidad de Codificación.

6. Electrónica y electricidad: Razonamiento Aritmético, Conocimiento Matemático, Información sobre Electrónica y Ciencia General.
7. Salud, sociedad y tecnología: Conocimiento de Palabras, Comprensión de Párrafos, Razonamiento Aritmético y Comprensión Mecánica.

El problema con la formación de combinaciones de esta manera es que se correlacionan en tal grado unas con otras que se vuelven esencialmente redundantes. De hecho, la intercorrelación promedio entre estas siete puntuaciones mixtas es de ¡.86! (Murphy, 1984). Es claro que las puntuaciones mixtas no siempre proporcionan información diferencial acerca de aptitudes específicas. Quizás ésta es la razón por la que las ediciones recientes de la ASVAB han evitado las puntuaciones mixtas múltiples y complejas. En vez de ello, el énfasis se ha colocado en puntuaciones mixtas más sencillas que se componen de constructos sumamente relacionados. Por ejemplo, la puntuación mixta de Capacidad Verbal se obtiene de Conocimiento de Palabras y Comprensión de Párrafos, dos subpruebas con elevada relación entre sí. De manera similar, la puntuación mixta en Capacidad Matemática se obtiene de la combinación de Razonamiento Aritmético y Conocimiento Matemático.

Algunos investigadores han concluido que la ASVAB no funciona como una batería de prueba de aptitudes múltiples, pero logra tener éxito en predecir diversas tareas militares debido a que las puntuaciones mixtas detectan, invariablemente, un factor general de inteligencia. La ASVAB puede ser una muy buena prueba de inteligencia general, pero no logra ser una batería de aptitudes múltiples. Otra preocupación es que la prueba puede poseer diferentes estructuras psicométricas para varones y mujeres. Específicamente, la subprueba de Información sobre Electrónica es una buena medida de g (el factor general de inteligencia) en el caso de los varones, pero no en el de las mujeres (Ree y Carretta, 1995). La posible explicación de esto es que aquéllos tienen una probabilidad nueve veces mayor de inscribirse en clases de electrónica y taller automotor en la secundaria y, por tanto, tienen la oportunidad de que su capacidad general moldee aquello que han aprendido, mientras las mujeres no la tienen. Así, las calificaciones en estas subpruebas funcionarán como una medida del rendimiento (aquello que ya se ha aprendido), pero no como un índice de aptitud (predicción de resultados futuros).

Desde el decenio de 1980 ha estado en proceso una investigación sobre una Prueba Adaptativa Computarizada (CAT), que es una versión de la ASVAB. Esta prueba se analiza en el tema 15A, Evaluación por computadora y el futuro de las pruebas. Aquí se proporciona un resumen. En la CAT, la persona realiza la prueba mientras está sentada en una terminal de computadora. El nivel de dificultad de los reactivos que aparecen en pantalla se reajusta de manera continua en función del desempeño que va presentando el individuo examinado. En general, una persona que responde bien a un reactivo de subprueba, recibirá uno más difícil, mientras que alguien que no lo hizo recibirá otro más sencillo. La computadora utiliza la teoría de respuesta al reactivo como una base para la selección de reactivos. Cada individuo examinado recibe un conjunto único de reactivos de prueba diseñado para su nivel de capacidad.

La CAT-ASVAB se sometió a prueba en un limitado número de lugares. Larson (1994) detalla las razones para adoptar la CAT-ASVAB:

1. Reduce el tiempo total de la prueba (las pruebas ajustables requieren aproximadamente la mitad de los reactivos que aquéllas estándar).
2. Aumenta la seguridad de la prueba al eliminar la posibilidad de que alguien robe los protocolos de la misma.
3. Aumenta la precisión de prueba en los extremos superior e inferior de capacidad.
4. Proporciona un medio para realimentación inmediata de las calificaciones de la prueba, dado que las computadoras utilizadas para el examen pueden calificar de inmediato la prueba y dar los resultados.
5. Proporcionan un medio para flexibilizar los tiempos de inicio de la prueba (a diferencia de las pruebas de lápiz y papel aplicadas de manera grupal, donde todos deben comenzar

y terminar en el mismo momento, las pruebas basadas en computadora pueden diseñarse para los horarios personales de los individuos evaluados).

El éxito o fracaso de la CAT-ASVAB puede tener implicaciones trascendentales para las pruebas en general. Si el proyecto tiene éxito, es probable que la ASVAB y otras pruebas se amplíen a fin de medir nuevos aspectos del desempeño, como las latencias de respuesta, y que presenten tipos únicos de reactivos como pruebas visoespaciales de objetos en movimiento (Larson, 1994). La CAT-ASVAB tiene el potencial de cambiar el futuro de las pruebas.

PREDICCIÓN DEL DESEMPEÑO EN LA UNIVERSIDAD

Como sabe casi cualquier estudiante universitario, uno de los principales usos de las pruebas de aptitud es la predicción del desempeño académico. En la mayor parte de los casos, los aspirantes a una universidad deben enfrentarse con los programas de evaluación de las *Scholastic Assessment Test*, SAT (Pruebas de Evaluación Escolar) o la *American College Test*, ACT (Prueba Universitaria Estadounidense). Es posible que las instituciones establezcan normas mínimas de admisión para las pruebas SAT o ACT, con base en el conocimiento de que las bajas calificaciones predicen el fracaso en la universidad. En esta sección se analizará la idoneidad técnica y la validez predictiva de las principales pruebas de aptitud para la universidad.

PRUEBAS DE EVALUACIÓN ESCOLAR

Conocida antes como *Scholastic Aptitude Test*, SAT (Pruebas de Aptitud Escolar), la *Scholastic Assessment Test*, SAT (Pruebas de Evaluación Escolar) son las más antiguas de las pruebas de admisión a la universidad, ya que datan de 1926. La Junta Universitaria (antes conocida como Junta de Examen de Ingreso a la Universidad), un grupo formado en 1899, publicó las SAT para establecer un centro nacional de información para las pruebas de admisión. Como lo señaló el historiador Fuess (1950), el propósito de una prueba de admisión a nivel nacional era "introducir orden y concierto en una anarquía educativa que hacia finales del siglo XIX se había vuelto exasperante, de hecho casi intolerable, para los catedráticos". A través de los años, la prueba se ha revisado de manera extensa, con actualizaciones continuas y renormatizaciones repetidas. A principios del decenio de 1990, se cambió el nombre de las SAT por el de *Scholastic Assessment Test* para enfatizar los cambios en contenido y formato. La nueva SAT evalúa el dominio en materias de educación media en mayor grado que su antecesora, pero continúa dirigiéndose a las habilidades de razonamiento. La SAT representa la vanguardia de las pruebas de aptitud.

La nueva SAT consta de las Pruebas de Razonamiento SAT-I y las Pruebas Temáticas SAT-II. La Prueba de Razonamiento Verbal de la SAT-I enfatiza vocabulario dentro de un contexto, comprensión de lectura y razonamiento crítico. Razonamiento Matemático de la SAT-I enfatiza la aplicación de conceptos matemáticos, interpretación de datos y construcción real de una respuesta, en contraposición al formato típico de opción múltiple. Se recomienda mucho el uso de una calculadora, pero esto no es obligatorio.

Las calificaciones de Razonamiento Verbal y Razonamiento Matemático de la SAT-I se presentan en una escala que va de 200 a 800 puntos. Los tipos de reactivos característicos de la sección Verbal incluyen:

- Analogías: seleccionar un par de palabras que mejor representen una relación similar a la expresada en un par de estímulos.
- Frases Incompletas: para una oración con 1 o 2 espacios en blanco, elegir una palabra o par de palabras que se ajusten mejor al significado global de la oración.
- Comprensión de Lectura: leer una selección y responder a preguntas de opción múltiple basadas en lo que se expresó en la selección de lectura.

Los tipos de reactivos característicos de la sección de Matemáticas incluyen los siguientes:

- Matemáticas Regulares: resolver problemas básicos de geometría y álgebra.
- Comparaciones Cuantitativas: elegir entre dos cantidades la mayor o indicar que son iguales o que el problema no puede resolverse con la información proporcionada.

Una concepción errónea persistente acerca de las calificaciones de la SAT es que 500 y 100 representan la media y desviación estándar de la muestra más reciente de personas examinadas con la prueba (Donlon, 1984). De hecho, los números 500 y 100 se refieren a la media y desviación estándar del grupo de referencia de 10 654 estudiantes que contestaron la sección Verbal de la SAT en abril de 1941. (La sección de Matemáticas se equiparó con esta sección al año siguiente.) Todas las nuevas calificaciones se equiparan con aquéllas de referencia al vincular cada nueva forma de la SAT con una o más formas anteriores. Por ejemplo, si una nueva forma es un tanto más fácil que las anteriores, puede ser necesario que la persona que realiza la prueba necesite algunas respuestas correctas adicionales a fin de lograr una calificación equivalente. Este procedimiento garantiza que las calificaciones actuales de la SAT se basen en la misma escala de medición utilizada al inicio del procedimiento de referencia en 1941.* En 1996 se terminó de establecer la nueva escala y se restablecieron las calificaciones de la SAT. Uno de los propósitos de la nueva escala era proporcionar mediciones más confiables en los rangos de calificación superior e inferior al ampliar el nivel de dificultad de los reactivos (Johnson, 1994).

De un año a otro, la calificación promedio de las personas evaluadas con la SAT puede ser sustancialmente diferente del promedio original de 500. De hecho, las calificaciones de la SAT disminuyeron de manera abrupta de 1963 a 1980. Para 1980, el promedio en Matemáticas había disminuido de 500 a cerca de 465, mientras que el promedio Verbal llegó hasta 420, casi una desviación estándar completa con respecto al punto inicial. Las calificaciones promedio en ambas escalas han aumentado sólo ligeramente desde entonces. Este fenómeno ha estado sujeto a intenso escrutinio, y revisar todas las explicaciones que se han proferido está más allá del alcance de este libro. En general, se aceptan las siguientes conclusiones:

- La disminución no fue un efecto artificial de la dificultad o de la escala de calificaciones de la SAT, fue un fenómeno real que afectó a otros de los principales programas de prueba.
- La disminución fue significativa y constituyó un cambio considerable en el desempeño en la prueba; el cambio representó un "grave deterioro del proceso de aprendizaje en EUA" (Wirtz, 1977).
- La disminución no redujo la validez predictiva de la SAT; la prueba mantuvo correlaciones adecuadas con el desempeño universitario.
- Los cambios poblacionales como el aumento en el tamaño de las familias pueden explicar parte de la disminución en calificaciones; si el tamaño promedio de las familias continúa reduciéndose, se predice que las calificaciones de la SAT aumentarán (Zajonc, 1976; Zajonc, Markus y Markus, 1979).
- Los cambios sociales como la expansión de la televisión pueden haber contribuido a la disminución en calificaciones; no obstante, tales hipótesis son difíciles de probar (Donlon, 1984).

Se ha cuidado la construcción de las nuevas formas de las pruebas Verbal y de Matemáticas de la SAT, debido a que una confiabilidad a toda prueba y un alto grado de paralelismo son esenciales para la misión de este programa de pruebas. La confiabilidad de consistencia interna de ambas formas se encuentra de manera repetida

* Esta práctica difiere del procedimiento que se siguió en la renormatización de las pruebas de CI. Cuando una revisión de una prueba de CI se estandariza de nuevo, la media de la población siempre se establece en 100, sin considerar la dificultad comparativa de la prueba revisada y de su antecesora.

en el rango de .91 a .93; con algunas excepciones, las correlaciones *test-retest* varían entre .87 y .89. El error estándar de medición es de 30 a 35 puntos.

La principal evidencia de la validez de la SAT se relaciona con el criterio, en este caso la capacidad para predecir las calificaciones en el primer año de universidad. Donlon (1984, capítulo VIII) presenta un cúmulo de información acerca de este punto; en este caso sólo pueden resumirse las tendencias. En 685 estudios, las calificaciones combinadas Verbal y de Matemáticas de la SAT se correlacionaron, en promedio, en .42 con el promedio de calificaciones del primer año en la universidad. Es interesante que el historial en educación media (p. ej., el rango o promedio de calificaciones) se desempeña mejor que la SAT en la predicción de las calificaciones universitarias ($r = .48$). Pero la combinación de la SAT y el historial en bachillerato resulta todavía mejor para la predicción; estas variables se correlacionan, en promedio, en .55 con el promedio de calificaciones del primer año de universidad. Por supuesto, estos datos reflejan una restricción sustancial de los límites: los estudiantes de bachillerato con bajas calificaciones en la SAT no asisten a la universidad. Donlon (1984) estima que la correlación real sin restricción de límites (SAT + calificaciones del bachillerato) se encontraría en las cercanías de .65.

Una cuestión de gran preocupación es el efecto del estudio y del entrenamiento especial sobre las calificaciones de la SAT. ¿Sirve de algo recibir entrenamiento especial en vocabulario o matemáticas o leer las numerosas guías de estudio que están disponibles en casi cualquier librería? Messick y Jungeblut (1981) revisaron los estudios disponibles que utilizaban un formato de grupo experimental contra grupo control. Estos investigadores concluyeron que el entrenamiento eleva las calificaciones combinadas de las pruebas Verbal y de Matemáticas en aproximadamente 28 a 30 puntos, lo cual no representa un incremento sustancial en comparación con la falta de entrenamiento/preparación. No obstante, para los estudiantes con alto nivel de motivación que buscan entrenamiento y reciben un programa estructurado, riguroso, los efectos del entrenamiento son mucho más grandes, de 45 a 110 puntos en las puntuaciones Verbal y de Matemáticas combinadas (Johnson, 1994). Una cuestión relacionada tiene que ver con la elevada proporción de estudiantes que contestan la SAT más de una vez. ¿Las calificaciones tienden a elevarse con la repetición de la prueba? En casos donde la repetición ocurre dentro de los siguientes 5 a 8 meses, el incremento promedio es de cerca de 12 puntos para cada una de las calificaciones Verbal y de Matemáticas (Donlon, 1984). Esto refleja, en parte, la familiaridad con la prueba, pero un factor que muchos pasan por alto es el aprendizaje activo que puede ocurrir en el ínterin.

PRUEBA UNIVERSITARIA ESTADOUNIDENSE

El programa de evaluación de la *American College Test*, ACT (Prueba Universitaria Estadounidense), es un sistema reciente de examen e informe diseñado para estudiantes que pretenden ingresar a la universidad. Además de las calificaciones tradicionales de la prueba, el programa de evaluación de la ACT incluye un inventario de intereses con 90 reactivos (basado en la tipología de Holland) y una sección sobre el perfil del estudiante (en la que el alumno puede listar las materias estudiadas, logros notables, experiencia laboral y servicios comunitarios). Aquí no se analizarán estas medidas complementarias, excepto para señalar que son útiles para generar el Informe del perfil del estudiante, que se envía a la persona examinada y a las universidades listadas en la carpeta de registro.

Iniciada en 1959, la ACT se basó en la filosofía de que las pruebas directas de las habilidades necesarias para los cursos universitarios proporcionan la base más eficaz para predecir el desempeño en la universidad. En términos del número de estudiantes que contestan la prueba, la ACT ocupa el segundo lugar después de la SAT como examen de admisión a la universidad. Las cuatro pruebas de la ACT requieren

del conocimiento en un área temática, pero enfatizan la utilización de ese conocimiento:

- Inglés (75 preguntas, 45 minutos). Se presentan al examinando varios fragmentos de prosa extraídos de escritos publicados. Ciertas partes del texto se encuentran subrayadas y numeradas, y se presentan posibles revisiones para las secciones señaladas; además existe una opción de "sin cambio". La persona debe elegir la mejor alternativa.
- Matemáticas (60 preguntas, 60 minutos). Aquí se le pide al estudiante evaluado que resuelva los tipos de problemas matemáticos que es probable que se presenten en los cursos básicos de matemáticas en la universidad. La prueba enfatiza los conceptos más que las fórmulas y utiliza un formato de opción múltiple.
- Lectura (40 preguntas, 35 minutos). Esta subprueba está diseñada para evaluar el nivel de la persona examinada en cuanto a comprensión de lectura; se presentan subpuntuaciones para las habilidades de lectura en ciencias sociales y artes/literatura.
- Razonamiento en ciencias naturales (40 preguntas, 35 minutos). Esta prueba evalúa la capacidad de lectura y comprensión de material de ciencias naturales. Las preguntas se extraen de representaciones de datos, resúmenes de investigación y puntos de vista conflictivos.

Además de las calificaciones de área que se señalan antes, los resultados de la ACT también se presentan como una Puntuación mixta total, que es el promedio de las cuatro pruebas. La calificación de la ACT se muestra en una escala de puntuación estándar de 36 puntos. En 1992, el promedio de la Puntuación mixta de la ACT para los graduados de bachillerato era de 20.7, con una desviación estándar de aproximadamente cinco puntos (Maxey, 1994). No obstante, como en la SAT, las calificaciones de la ACT no son fijas para algún año en particular; mostraron la misma disminución en los decenios de 1960 y 1970 que se observó en la SAT.

Los críticos del programa ACT han señalado el énfasis en la comprensión de lectura que satura las cuatro pruebas. La intercorrelación promedio de las pruebas se encuentra por lo común alrededor de .60. Estos datos sugieren que un factor general de rendimiento/capacidad domina las cuatro pruebas; por tanto, no deben sobreinterpretarse los resultados de cualquier prueba específica. Por fortuna, es probable que las personas encargadas de la admisión a la universidad le den mayor énfasis a la Puntuación mixta, que es el promedio de las cuatro pruebas independientes. La ACT parece medir, en gran parte, lo mismo que la SAT; la correlación entre ambas se acerca a .90. Por tanto, no es de sorprender que la validez predictiva de la Puntuación mixta de la ACT rivalice con la de la puntuación combinada de la SAT, con correlaciones en las cercanías de .40 y .50 entre la puntuación mixta de la ACT y el promedio de calificaciones del primer año de universidad. Los coeficientes de validez predictiva son casi idénticos para estudiantes de todo tipo de ambientes y circunstancias, lo cual indica que las pruebas ACT son imparciales.

Kifer (1985) no cuestiona la idoneidad técnica de la ACT y de programas similares de prueba, pero sí protesta contra el enorme poder simbólico que han acumulado estos exámenes. El énfasis en las calificaciones de la prueba para las admisiones a la universidad no es una cuestión ética, sino una preocupación social, moral y política:

> Las admisiones selectivas implican simplemente que una institución no podrá o no querrá admitir a cualquier persona que llene una solicitud. La decisión de a quién se admitirá y a quién no, debería ser, en primer lugar, una cuestión de aquello que la institución considera deseable y puede o no incluir el uso de ecuaciones de predicción. Es tan justificable seleccionar con base en el talento, interpretado de manera amplia, como lo es utilizar las calificaciones de una prueba, por altas que sean. Existen estudiantes talentosos en muchas áreas —líderes, organizadores, creadores, músicos, atletas, ganadores de premios científicos, aficionados a la ópera— que pueden tener calificaciones moderadas o bajas en la ACT, pero cuya presencia en una universidad la transformarían.

Es posible que el lector desee revisar el tema 7B (Sesgo de una prueba y otras controversias) para un mayor análisis sobre este punto.

PRUEBAS DE SELECCIÓN DE POSGRADOS

Los programas de posgrado y profesionales también dependen, en gran medida, de las pruebas de aptitud para las decisiones de admisión. Por supuesto, se toman en consideración muchos otros factores cuando se selecciona a estudiantes para entrenamiento avanzado, pero no se puede negar la función central de los resultados de las pruebas de aptitud para las decisiones de selección. Por ejemplo, la figura 8–5 presenta un sistema común de importancia cuantitativa que se utiliza en la evaluación de solicitantes para el entrenamiento de posgrado en psicología. El lector verá que la calificación total en el *Graduate Record Exam*, GRE (Examen de Selección para Posgrado), recibe la mayor importancia en el proceso de selección. En las siguientes secciones se revisa el GRE, así como las pruebas de admisión que se utilizan en las escuelas de medicina y leyes.

EXAMEN DE SELECCIÓN PARA POSGRADO

El GRE es una prueba grupal de opción múltiple que se utiliza de manera amplia en los programas de posgrado en muchas áreas como un componente en la selección de candidatos para entrenamiento avanzado; ofrece exámenes temáticos en muchos campos (p. ej., Biología, Ciencias Informáticas, Historia, Matemáticas, Ciencias Políticas, Psicología), pero el alma de la evaluación se encuentra en la prueba general que está diseñada para medir las aptitudes verbales, cuantitativas y analíticas. La sección verbal (GRE-V) incluye reactivos verbales como analogías, terminación de oraciones, antónimos y comprensión de lectura. La sección cuantitativa (GRE-Q) consiste en problemas de álgebra, geometría, razonamiento y la interpretación de datos, gráficas y diagramas. La sección analítica (GRE-A) incluye problemas de razonamiento analítico y lógico. Las tres calificaciones del GRE se presentan como puntuaciones estándar con una media **aproximada** de 500 y desviación estándar de 100. De hecho, la calificación media puede diferir de un año a otro, debido a que todos los resultados de la prueba se vinculan con un grupo estándar de referencia de 2 095 estudiantes de los últimos años de la universidad evaluados en 1952 en las partes verbal y cuantitativa de la prueba. Los programas de posgrado tienden a prestar atención a la combinación de calificaciones en las primeras dos partes (GRE-V + GRE-Q), donde las califica-

Calificaciones GRE total GRE-V + GRE-Q total:	0	6 1 000	12 1 100	18 1 200	24 1 300	30 1 400
Promedio general de calificaciones (PGC) en pregrado	0	5 3.0	10 3.2	15 3.4	20 3.6	25 3.8
PGC en Psicología	0	1 3.0	2 3.2	3 3.4	4 3.6	5 3.9
Antecedentes en estadística/experimentales	0	1	2	3	4	5
Antecedentes en Biología/Química	0	1	2	3	4	5
Antecedentes en Matemáticas/Ciencias informáticas	0	1	2	3	4	5
Experiencia en investigación	0	1	2	3	4	5
Habilidades interpersonales positivas	0	2	4	6	8	10
Diversidad étnica/lingüística/ cultural	0	2	4	6	8	10
						Total máximo: 100

Figura 8–5. Esquema de ponderaciones representativas utilizado por los Comités de Admisión al Programa de Posgrado en Psicología.

ciones combinadas que superen a 1 000 se considerarían superiores al promedio.

La confiabilidad del GRE es fuerte, con coeficientes de confiabilidad de consistencia interna que por lo común se encuentran alrededor de .90 para los tres componentes. Por lo general, la validez del GRE se ha examinado en relación con la capacidad de la prueba para predecir el desempeño en la escuela de posgrado. El desempeño se ha medido principalmente de acuerdo con el promedio de calificaciones, aunque también se han utilizado las calificaciones de los académicos sobre la aptitud del estudiante. Por ejemplo, con base en una revisión metaanalítica de 22 estudios con un total de 5 186 estudiantes, Morrison y Morrison (1995) concluyeron que el GRE-V se correlacionaba en .28 y el GRE-Q en .22 con el promedio de calificaciones en posgrado. Así, en promedio, las calificaciones en el GRE explicaban sólo 6.3% de la varianza en el desempeño académico a nivel de posgrado. En un estudio reciente con 170 estudiantes de posgrado en psicología de la Universidad de Yale, Sternberg y Williams (1997) también encontraron correlaciones mínimas entre las puntuaciones del GRE y las calificaciones del posgrado. Cuando se correlacionaron las puntuaciones en el GRE y las calificaciones de los académicos en cinco variables (capacidades analíticas, creativas, prácticas, de investigación y de enseñanza), las correlaciones fueron incluso menores, donde la mayor parte oscilaba alrededor de cero. La única excepción fue la puntuación en el GRE-V, que se correlacionó poco con casi todas las calificaciones de los profesores. No obstante, esta correlación se observó **sólo** en el caso de los varones (en el orden de $r = .3$), ¡mientras para las mujeres fue exactamente de cero en todos los casos! Con base en éstos y otros estudios similares, el consenso parecería ser que la dependencia excesiva en el GRE para la selección a la escuela de posgrado podría pasar por alto una reserva de talentos con prometedores estudiantes de posgrado.

A pesar de ello, otros investigadores son más optimistas en su valoración del GRE, al señalar que la correlación entre las puntuaciones de la prueba y las calificaciones del posgrado no es un buen índice de validez debido al problema de la restricción del rango (Kuncell, Campbell y Ones, 1998). Específicamente, los solicitantes con bajas calificaciones en el GRE tienen menos probabilidades de que se les acepte de inicio para el entrenamiento en posgrado y, por tanto, existe relativamente poca información disponible con respecto a si las bajas calificaciones predicen un desempeño académico deficiente. Dicho de manera sencilla, la correlación de las calificaciones del GRE con el desempeño académico en posgrado se basa **principalmente** en personas con calificaciones en los niveles medio a alto en esta prueba, es decir, con totales GRE-V + GRE-Q de 1 000 o más. Como tal, la correlación se atenuaría precisamente debido a que no se incluye en la muestra a aquellos que tienen bajas calificaciones en el GRE. Otro problema con la validación de esta prueba contra las calificaciones en la escuela de posgrado es la falta de confiabilidad del criterio (calificaciones). Con base en la expectativa de que los estudiantes de posgrado tendrán un desempeño a niveles elevados, algunos profesores pueden dar dieces indiscriminados, de modo que las calificaciones no reflejan diferencias reales en las aptitudes de los estudiantes. Esto reduciría la correlación entre el medio de predicción (puntuaciones en el GRE) y el criterio (calificaciones del posgrado). Cuando se aclaran estos factores, muchos investigadores encuentran razones para considerar que el GRE aún es una herramienta válida para la selección en posgrado (Melchert, 1998; Ruscio, 1998).

PRUEBA DE ADMISIÓN A LA FACULTAD DE MEDICINA

La *Medical College Admission Test*, MCAT (Prueba de Admisión a la Facultad de Medicina) es un requisito para los solicitantes de casi todas las escuelas de medicina en EUA. La prueba se diseñó para evaluar el rendimiento en las habilidades y conceptos básicos que son requisito previo para la terminación exitosa de los

estudios de medicina. Tiene tres secciones de opción múltiple (Razonamiento Verbal, Ciencias Físicas y Ciencias Biológicas) y una sección de ensayo (Muestra de Redacción). La sección de Razonamiento Verbal se diseñó para evaluar la capacidad para comprender y aplicar la información y argumentos presentados en forma escrita. La prueba consiste, específicamente, en varios pasajes de aproximadamente 500 a 600 palabras cada uno, tomados de trabajos sobre humanidades, ciencias sociales y ciencias naturales. A cada uno le siguen varias preguntas basadas en la información que se incluye en el texto. La sección de Ciencias Físicas está diseñada para evaluar el razonamiento en química y física generales. Ciencias Biológicas se diseñó para valorar el razonamiento en biología y química orgánica. Estas secciones de ciencias físicas y biológicas contienen de 10 a 11 series de problemas descritos en cerca de 250 palabras cada uno, a los que les siguen varias preguntas. La prueba de Muestra de Redacción consiste en dos ensayos de 30 minutos. Esta prueba se diseñó para evaluar las habilidades básicas de redacción como desarrollo de una idea central, síntesis de conceptos e ideas, redactar de manera lógica y seguimiento de las prácticas aceptadas en cuanto a gramática, sintaxis y puntuación.

Cada una de las calificaciones de la MCAT se presenta en una escala de 1 a 15 (medias de aproximadamente 8.0 y desviación estándar de cerca de 2.5). La confiabilidad de la prueba es menor que la de otras pruebas de aptitud que se utilizan para selección, con coeficientes de consistencia interna y por mitades que principalmente se encuentran en la parte baja del rango de .80 (Gregory, 1994a). Las calificaciones de la MCAT predicen de manera mínima el éxito en la escuela de medicina, pero de nuevo esto se debe a la participación del enigma de la restricción del rango (que se analizó anteriormente en relación con el GRE). En particular, en pocas ocasiones se admite a los individuos examinados que tienen bajas calificaciones en la MCAT y que presumiblemente confirmarían la validez de la prueba al tener un desempeño deficiente en la escuela de medicina, lo cual reduce la validez aparente de la prueba.

PRUEBA DE ADMISIÓN A LA FACULTAD DE LEYES

La *Law School Admission Test*, LSAT (Prueba de Admisión a la Facultad de Leyes), es una prueba estandarizada que se aplica en medio día y que es requisito para los solicitantes de casi todas las escuelas de leyes en EUA. La prueba se diseñó para medir las habilidades que se consideran esenciales para tener éxito en la escuela de leyes, incluyendo lectura y comprensión de material complejo, organización y manejo de información y capacidad para razonar de manera crítica y obtener inferencias correctas. La LSAT consta de preguntas de opción múltiple en cuatro áreas: comprensión de lectura, razonamiento analítico y dos secciones de razonamiento lógico. Se utiliza una sección adicional para someter nuevos reactivos a una prueba preliminar y realizar una adecuación previa de nuevas formas de prueba, pero esta sección no contribuye a la calificación de la LSAT. La escala de calificaciones para la LSAT se extiende desde un nivel bajo de 120 hasta uno alto de 180. Además de las partes objetivas, se aplica una muestra de redacción de 30 minutos al final de la prueba. La sección no se califica, pero se envían copias de esta muestra a todas las escuelas de leyes a las que hizo solicitud la persona evaluada.

La LSAT tiene confiabilidad aceptable (coeficientes de consistencia interna a la mitad del rango de .90) y se le considera como un medio de predicción moderadamente válido de las calificaciones en la escuela de leyes. No obstante, en un fascinante estudio, las calificaciones de la LSAT se correlacionaron en mayor medida con los resultados del examen de la barra de abogados estatal que con las calificaciones de la escuela de leyes (Melton, 1985). Esto habla bien de la validez de la prueba, en la medida en que vincula las calificaciones de esta prueba con un criterio importante y de la vida real.

RESUMEN

1. Las pruebas de aptitud y capacidad difieren en enfoque y utilización. En general, la capacidad es un concepto amplio, mientras la aptitud se refiere, por lo común, a segmentos homogéneos de capacidad. También, las pruebas de capacidad intentan estimar el funcionamiento actual, mientras las de aptitud suelen emplearse para ayudar a predecir el desempeño futuro.
2. Las pruebas de aptitud deben sus orígenes al análisis factorial, una familia de procedimientos empleados para resumir las relaciones entre variables que se correlacionan de maneras sumamente complejas. Por ejemplo, el análisis factorial podría ayudar a un investigador a descubrir que una batería de 24 pruebas de capacidad representa sólo cuatro variables subyacentes, llamadas factores.
3. El punto inicial de todo análisis factorial es la matriz de correlación, un cuadro completo de intercorrelaciones entre todas las variables; estas últimas pueden incluir resultados en cualquier dimensión más o menos continua; por ejemplo, calificaciones de la prueba, clase social y calificaciones conductuales.
4. La matriz factorial consiste en un cuadro de cargas factoriales que muestran la importancia de cada variable en cada factor. Un factor es una suma lineal ponderada de las variables. La carga factorial para cada variable es un coeficiente de correlación entre el factor y esa variable.
5. Los factores pueden trazarse como ejes geométricos de referencia, y las cargas para cada variable en cada factor pueden representarse dentro de ese espacio; esto permite que el investigador visualice la localización de cada variable en los 2 o 3 factores más importantes.
6. Dado que la posición de los ejes de referencia es arbitraria, el investigador está en libertad de girar los ejes de modo que produzcan la correspondencia más sensata con las cargas factoriales para las variables. Existen diversos métodos de rotación (p. ej., rotación hacia variedad positiva y rotación hacia estructura simple).
7. El nombre que se da a los factores requiere juicio e inferencia. En particular, el investigador debe hacer el intento de determinar los procesos y capacidades comunes que comparten las pruebas o variables con fuertes cargas en un factor. También, las pruebas o variables con bajas cargas pueden ayudar a afinar la definición y denominación de éste.
8. Para que pueda surgir un tipo particular de factor a partir de un análisis, algunas de las pruebas y medidas deben contener de inicio ese factor. Se prefieren muestras grandes, mayores a 200 personas. Es importante la elección de estrategias para la rotación: los ejes ortogonales suponen que los factores no tienen correlación, mientras los ejes oblicuos suponen que los factores están correlacionados.
9. El (DAT) que se utiliza para la orientación educativa y vocacional en los años de educación media, consta de ocho pruebas independientes. Una preocupación acerca de esta respetada batería es la falta de validez discriminativa entre las ocho pruebas..
10. La GATB es típica de las baterías de aptitudes múltiples basadas en el análisis factorial. Consta de ocho pruebas de lápiz y papel, y cuatro medidas con aparatos derivadas de un análisis factorial de 59 pruebas. La batería produce nueve puntuaciones factoriales útiles en la predicción del desempeño en el trabajo.
11. La ASVAB, que se utiliza en las fuerzas armadas para seleccionar y asignar a los reclutas, es probablemente la prueba de lápiz y papel de uso más amplio que se tenga en existencia. De la prueba se obtienen puntuaciones mixtas, cada una basada en las

calificaciones de 2 a 4 subpruebas. Una limitación de la ASVAB es que las puntuaciones mixtas tienen altas correlaciones entre sí.

12. La predicción del desempeño académico se facilita con pruebas como las SAT, y el programa de evaluación de la ACT.
13. La SAT produce calificaciones independientes en las áreas Verbal y de Matemáticas en referencia a la media de 500 (DE de 100) obtenida en 1941. La ACT produce cuatro calificaciones (inglés, matemáticas, lectura y razonamiento en ciencias naturales) que se presentan en una escala de puntuación estándar de 36 puntos (media de aproximadamente 20.7 y DE de 5).
14. El GRE, que se utiliza para admisión a muchos programas de posgrado, consiste en tres pruebas generales (Verbal, Cuantitativa y Analítica), así como de pruebas temáticas en áreas específicas (p. ej., Biología, Ciencias Informáticas y Psicología). Las normas de cada examen se establecieron con una media aproximada de 500 y DE de 100.
15. La MCAT es requisito para los solicitantes de casi todas las escuelas de medicina en EUA. Tiene tres secciones de opción múltiple (Razonamiento Verbal, Ciencias Físicas y Ciencias Biológicas) y una sección de ensayo (Muestra de Redacción), todas diseñadas para evaluar las habilidades básicas necesarias para la escuela de medicina.
16. La LSAT es una prueba estandarizada que se aplica en medio día y que es requisito para los solicitantes de casi todas las escuelas de leyes en EUA. Consiste en preguntas de opción múltiple en cuatro áreas: Comprensión de Lectura, Razonamiento Analítico y dos secciones de Razonamiento Lógico.

TÉRMINOS Y CONCEPTOS CLAVE

Análisis factorial p. 322
Factores p. 322
Matriz de correlación p. 323
Matriz factorial p. 325
Rotación hacia variedad positiva p. 326
Rotación hacia estructura simple p. 326
Ejes ortogonales p. 330
Ejes oblicuos p. 330

TEMA 8B
Pruebas grupales de rendimiento

Conceptos esenciales en pruebas de rendimiento
Pruebas de rendimiento educativo
Pruebas de rendimiento con propósitos específicos
Falseamiento: el lado oscuro de las pruebas de rendimiento
Resumen
Términos y conceptos clave

En este tema se continúa con el análisis de las pruebas grupales al estudiar su empleo dentro de los sistemas educativos. Desde los años de la educación primaria, los distritos escolares utilizan pruebas grupales de rendimiento para seguir el progreso de los estudiantes individuales y estimar el éxito de los programas educativos. La difundida práctica de estas pruebas dentro de las escuelas estadounidenses es, en gran medida, un asunto positivo debido a que proporciona una base objetiva para la evaluación. No obstante, en ocasiones también tiene un lado oscuro, en la medida en que la prueba puede volverse en contra de las expectativas. El impacto negativo puede encontrarse en tres categorías generales. Primera, los maestros pueden enseñar para pasar la prueba, en vez de tratar de impartir un conocimiento genuino. Segunda, en la búsqueda por obtener altas calificaciones para sus sistemas escolares, los administradores pueden fomentar un ambiente que aliente las pruebas liberales, no estandarizadas. Peor aún, el personal de la escuela puede participar en un fraude explícito como "corregir" las hojas de respuesta. La tercera consecuencia es que los examinandos individuales encontrarán maneras ingeniosas de falsear los resultados en pruebas con normas nacionales. Al final de este tema se revisan algunas de estas tendencias alarmantes.

CONCEPTOS ESENCIALES EN PRUEBAS DE RENDIMIENTO

Las pruebas de rendimiento, conocidas como pruebas de logro en el Reino Unido, son las de uso más amplio entre todos los tipos de prueba. Aunque no existen cifras precisas en cuanto a utilización, en EUA casi cualquier niño en edad escolar se enfrenta anual o bianualmente a una prueba grupal estandarizada de rendimiento. Un estimado es que las escuelas públicas aplican un promedio de dos y media pruebas por estudiante cada año (Medina y Neill, 1990). Sin duda, el número de pruebas de rendimiento que se aplican sobrepasa todas las otras formas de evaluación psicológica y educativa.

Las pruebas de rendimiento están diseñadas para medir la obtención de las habilidades que se enseñan dentro de programas escolares o de entrenamiento. Estas pruebas pueden definirse de manera bastante estrecha, como en el caso de una prueba de habilidades de calificación, o concebirse de modo más amplio, como en el caso de una prueba de comprensión de lectura. Aunque las pruebas de rendimiento difieren en su especificidad, todas satisfacen una función relacionada: medir el nivel actual de habilidad en un área bien definida.

Como se les cataloga en la serie del *Mental Measurements Yearbook* (Conoley e Impara, 1995; Conoley y Kramer, 1989, 1992; Mitchell, 1985), se han publicado literalmente cientos de estas pruebas, lo que dificulta revisar el amplio panorama de estos instrumentos; en vez de ello, se revisarán las pruebas más representativas y enfocándose en los problemas que surgen debido a su empleo. Se comienza con un compendio acerca de los conceptos esenciales en el examen del rendimiento.

PRUEBAS GRUPALES E INDIVIDUALES DE RENDIMIENTO

Se ha establecido una distinción fundamental entre las pruebas grupales e individuales de rendimiento. Las primeras se utilizan principalmente en el salón de clase, mientras las segundas se usan en entornos clínicos y educativos donde la evaluación es personal. Las **pruebas grupales de rendimiento** podrían llamarse también pruebas de aprovechamiento educativo, ya que es común que estos instrumentos se apliquen a todo un sistema escolar a instancias de los superintendentes escolares u otros administradores dentro de una coordinación estatal. Por supuesto, las pruebas grupales se aplican de manera simultánea a docenas o cientos de estudiantes, con todas las ventajas y desventajas que este enfoque conlleva (tema 2B, Proceso de aplicación).

Las **pruebas individuales de rendimiento** representan una función esencial en el diagnóstico de la discapacidad para el aprendizaje (DA), pues no sólo proporcionan documentación para las alteraciones en el desempeño en áreas académicas importantes como la lectura, escritura y cálculo, sino también algunas de ellas pueden ayudar a identificar las deficiencias particulares en cuanto a habilidades que subyacen a las discapacidades para el aprendizaje. Como se analiza más adelante, las pruebas individuales de rendimiento se utilizan junto con otros instrumentos, en especial pruebas de inteligencia, como se analiza en el tema 10A Evaluación escolar.

PRUEBAS REFERIDAS A NORMAS Y A CRITERIO

Además de la dicotomía fundamental que divide a las pruebas de rendimiento en grupales e individuales, otra importante distinción es aquella entre las pruebas de rendimiento referidas a normas y a criterio. El lector recordará del tema 2A (Naturaleza y usos de las pruebas psicológicas) que las pruebas referidas a normas permiten la interpretación en relación con una gran muestra de estandarización. Estas pruebas facilitan el informe de las calificaciones como rangos percentiles, puntuaciones estándar y similares. En contraste, las pruebas referidas a criterio permiten la interpretación con relación al contenido específico que domina el individuo examinado. Por ejemplo, una prueba referida a criterio podría determinar que una persona

sabe la ortografía correcta de 94 de los 100 reactivos de una lista específica de palabras esenciales. Por supuesto, ambos enfoques no son necesariamente incompatibles. De hecho, la mayor parte de las principales pruebas de rendimiento proporcionan interpretaciones tanto referidas a normas como a criterio.

PRUEBAS DE CAPACIDAD, APTITUD Y RENDIMIENTO

La distinción entre pruebas de capacidad, aptitud y rendimiento merece una breve revisión dentro de este contexto. Las primeras toman una muestra de un amplio conjunto de habilidades a fin de estimar el nivel intelectual general. En contraste, las de aptitud suelen medir segmentos homogéneos de capacidad y con frecuencia se les utiliza para predecir el desempeño futuro. Las excepciones en este caso incluyen las baterías de prueba de aptitudes múltiples, que toman una muestra amplia de capacidades; estos instrumentos son muy similares a las pruebas de capacidad. Por último, como se señaló antes, las pruebas de rendimiento miden la obtención de habilidades en el momento presente, en particular con relación a los programas escolares y de entrenamiento.

En el mundo real, la distinción entre estos tres tipos de pruebas es, a menudo, bastante confusa (Gregory, 1994a). Durante algún tiempo se ha sabido que la correlación entre las pruebas de rendimiento y de capacidad puede ser casi tan elevada como la que existe entre dos pruebas de capacidad. En muchos casos, ambas pruebas tocan factores cognitivos subyacentes muy similares. No obstante, las suposiciones que se encuentran como base para estas dos formas de prueba difieren mucho. En general, las pruebas de rendimiento están diseñadas para medir los efectos de experiencias educativas relativamente estandarizadas, mientras que, por lo común, las pruebas de capacidad hacen pocas suposiciones acerca de las experiencias específicas de aprendizaje que hayan ocurrido con anterioridad.

Las aplicaciones de las pruebas de aptitud y de rendimiento también difieren en gran medida. Las primeras están diseñadas primordialmente para predecir el desempeño futuro en programas escolares o de entrenamiento. Por ejemplo, podría utilizarse una prueba de aptitud escolar para predecir el desempeño académico futuro en la universidad o emplearse una prueba de aptitud para trabajos de oficina, a fin de pronosticar el desempeño futuro en un puesto secretarial. En contraste, las pruebas de rendimiento se usan para estimar el nivel actual de logro de un estudiante en una materia específica. En otras palabras, las pruebas de aptitud están orientadas al futuro, mientras las de rendimiento se orientan al presente. La evaluación del nivel actual de habilidad a través de pruebas de rendimiento puede satisfacer varios propósitos, como se expresa a continuación.

FUNCIONES DE LAS PRUEBAS DE RENDIMIENTO

Estas pruebas tienen diversos usos potenciales. Las aplicaciones prácticas de las pruebas individuales y grupales de rendimiento incluyen las siguientes:

- Identificar a niños y adultos con deficiencias específicas en cuanto a rendimiento y que podrían necesitar de una evaluación más minuciosa de discapacidades para el aprendizaje.
- Ayudar a los padres a reconocer las aptitudes y debilidades académicas de sus hijos y, por tanto, fomentar los esfuerzos individuales de regularización en el hogar.
- Identificar las deficiencias en rendimiento a nivel de la clase y de la escuela como una base para redirigir los esfuerzos educativos.
- Valorar el éxito de los programas educativos al medir el logro subsecuente de habilidades de parte de los alumnos.
- Agrupar a los estudiantes de acuerdo con el nivel similar de habilidad en áreas académicas específicas.
- Identificar el nivel de instrucción apropiado para estudiantes individuales.

Así, las pruebas de rendimiento satisfacen metas institucionales como el registro de los

niveles de rendimiento a nivel escolar, pero también representan un importante papel en la evaluación de las dificultades individuales de aprendizaje. Como se señaló antes, se utilizan diferentes tipos de pruebas de rendimiento para lograr estas dos aplicaciones fundamentales (institucionales e individuales). Las metas institucionales se logran mejor a través de las baterías de pruebas grupales de rendimiento, mientras la evaluación individual se realiza mejor con pruebas individuales de rendimiento (aunque las primeras también pueden representar un papel en este contexto). En este tema se centra la atención en las pruebas grupales de rendimiento educativo.

PRUEBAS DE RENDIMIENTO EDUCATIVO

Casi todos los sistemas escolares en EUA emplean cuando menos una prueba de rendimiento educativo, de modo que no es de sorprender que los editores de pruebas hayan respondido a esta necesidad generalizada al desarrollar una colección de instrumentos excelentes. En la siguiente sección se describen varias de las pruebas estandarizadas de rendimiento para grupos que más se utilizan. Las pruebas a describir comparten varias características.

Primero, estos instrumentos son baterías multinivel que contienen subpruebas comparables para alumnos en los diferentes grados de escuela primaria, secundaria o ambas. Algunas de las baterías abarcan desde el jardín de niños (K) hasta bachillerato, mientras otras están diseñadas sólo para los niveles de escuela primaria (K hasta octavo) o de bachillerato (9 a 12). En una batería multinivel, los protocolos de prueba contienen secciones superpuestas y los estudiantes en los diferentes niveles, según el grado escolar, inician y terminan con los materiales de prueba en la posición apropiada para su grado.

Un segundo aspecto común en muchas de las baterías de pruebas educativas, es la normatización concurrente con una prueba de capacidad. Por ejemplo, las normas de la batería de rendimiento conocida como Pruebas secuenciales de progreso educativo *Sequential Test of Educational Progress* STEP-II (Pruebas Secuenciales de Progreso Educativo), se obtuvieron de manera concurrente con las normas de la batería de capacidad conocida como *School and College Ability Test*, SCAT-III (Prueba de Capacidad Escolar y Universitaria). Las pruebas cuyas normas se establecen de modo concurrente comparten la misma muestra de estandarización. Como resultado, el desempeño promedio en una prueba puede compararse de manera directa con el obtenido en la otra. Las normas concurrentes son útiles debido a que permiten que padres, maestros y orientadores hagan comparaciones precisas, directas y significativas entre rendimiento y capacidad. Después de todo, las implicaciones de una calificación de rendimiento se moderan con base en el conocimiento de cuál es la capacidad del estudiante. Aquél con altas calificaciones en capacidad, pero bajas en rendimiento, podría ser buen candidato para la intervención educativa, incluyendo una evaluación más detallada de una posible discapacidad para el aprendizaje (como se analiza en el tema 10A, Evaluación escolar). En contraste, un estudiante con bajas calificaciones en capacidad y en rendimiento podría estar trabajando a todo su potencial; en este caso es posible que las intervenciones especiales no se justifiquen.

La tercera característica común en las pruebas grupales de rendimiento es que miden habilidades educativas similares. Las pruebas de rendimiento educativo tienden a enfatizar las siguientes áreas de habilidad:

- Lectura, incluyendo comprensión y vocabulario.
- Lenguaje escrito, incluyendo ortografía, puntuación y uso de letras mayúsculas.
- Matemáticas, incluyendo cálculo y aplicaciones.

Además, las pruebas para los niveles de primaria evalúan con frecuencia habilidades auditivas, incluyendo comprensión oral. Algunas baterías de prueba también evalúan conocimiento de conceptos básicos en ciencias naturales, ciencias sociales y humanidades.

Por último, las pruebas de rendimiento educativo que se analizan aquí poseen, en general, características psicométricas excelentes. Los contenidos de la prueba son pertinentes y apropiados; es decir, los instrumentos muestran buena validez de contenido. Las subescalas poseen excelente confiabilidad interna y de formas alternas. Las muestras de estandarización son grandes y representativas invariablemente, y las pruebas no muestran sesgo abierto de género o racial. La calidad psicométrica de las pruebas de rendimiento educativo de uso más común es típicamente respetable, por no decir extraordinaria.

A continuación se examinan varias pruebas de rendimiento educativo de amplio uso. El lector descubrirá que un análisis detallado de las propiedades psicométricas —confiabilidad, validez, normas y similares— se encuentra sólo en el primer instrumento revisado, las *Iowa Tests of Basic Skills*, ITBS (Pruebas Iowa de Habilidades Básicas). En general, la calidad psicométrica de las otras pruebas es igualmente loable, de modo que en estas baterías de prueba nos enfocamos en sus funciones, aplicaciones, características especiales y en una o dos deficiencias ocasionales. Los lectores que deseen más información sobre estos instrumentos deben consultar las revisiones en el *Mental Measurements Yearbook* (Conoley e Impara, 1995; Conoley y Kramer, 1989, 1992; Mitchell, 1985).

PRUEBAS IOWA DE HABILIDADES BÁSICAS

Publicadas originalmente en 1935, las *Iowa Tests of Basic Skills*, ITBS (Pruebas Iowa de Habilidades Básicas), tuvieron su revisión y reestandarización más reciente en 1992. Las ITBS constituyen una batería multinivel de pruebas de rendimiento que abarcan los grados K-9. Una prueba complementaria, los *Tests of Achievement and Proficiency*, TAP (Pruebas Vocacional de Rendimiento), cubre los niveles de bachillerato. A fin de hacer más expeditas las comparaciones directas y precisas entre rendimiento y capacidad, las normas de la ITBS y de la TAP se obtuvieron al mismo tiempo que las de la *Cognitive Abilities Test*, CogAT, (Prueba de Capacidades Cognitivas), un respetado grupo de pruebas de capacidad intelectual general.

La ITBS se encuentra disponible en varios niveles que corresponden aproximadamente a las edades de los posibles individuos a evaluar: niveles 5 y 6 (grados K-1), niveles 7 a 9 (grados 1 a 3) y niveles 9 a 14 (grados 3 a 9). Las subpruebas básicas para los niveles de mayor edad miden vocabulario, lectura, lenguaje, matemáticas, ciencias sociales, ciencias naturales y fuentes de información.

Desde la primera edición, la ITBS se ha guiado por una filosofía práctica de la medición en educación. El manual expresa el propósito del método de prueba como sigue:

> El propósito de la medición consiste en proporcionar información que pueda emplearse para mejorar la enseñanza. La medición tiene valor en la medida que dé por resultado mejores decisiones que afecten de manera directa a los alumnos.

Para lograr esta meta, la ITBS incorpora un análisis de habilidades referidas a criterio a fin de complementar el conjunto habitual de calificaciones referidas a normas. Por ejemplo, una característica disponible de parte del servicio de calificación del editor es la información a nivel del reactivo. Ésta indica áreas temáticas, reactivos que muestrean el tema y respuesta correcta o incorrecta para cada uno. Por tanto, los maestros tienen acceso a un cúmulo de información diagnóstica-educativa para cada estudiante. Es muy difícil cuantificar si esta información se traduce en una mejor enseñanza, como lo desean sus autores. Como señala Linn (1989), "debemos depender en gran medida de la lógica, las anécdotas y las opiniones cuando se trata de responder a tales cuestiones".

Las propiedades técnicas de la ITBS están más allá de dudas. Los coeficientes de confiabilidad de consistencia interna y de formas equivalentes se encuentran en su mayor parte en la mitad del rango de .80 y a principios del de .90. Los coeficientes de estabilidad para un intervalo de un año se encuentran casi todos en el rango entre .70 y .90. La prueba está libre de sesgo explícito de género y racial, como lo determinan

la evaluación del contenido y los estudios sobre sesgo de reactivos. La muestra de estandarización de la última edición incluyó a 270 000 estudiantes estratificados aproximadamente según el tipo de escuela (pública/privada/parroquial), ciudades grandes y pequeñas, regiones geográficas y niveles socioeconómicos.

La estandarización de una forma anterior en 1988 reveló una interesante tendencia en comparación con los resultados para versiones que se estandarizaron varios años antes. La muestra de 1988 demostró niveles más altos de rendimiento, en el orden de 1 a 3 meses de equivalencia de grado. Este patrón de elevación lenta del desempeño en una prueba enfatiza la necesidad de reestandarización anual o bienal de las baterías de pruebas de rendimiento. Lo que ha sucedido en ausencia de reestandarizaciones oportunas de las principales pruebas de rendimiento es que los 50 estados de EUA pueden informar con toda honradez que superan el promedio nacional en pruebas grupales estandarizadas (Cannell, 1988).

Expertos en el currículo de materias y críticos de la prueba consideran que los reactivos de la ITBS son pertinentes, lo cual habla en favor de su validez de contenido (Lane, 1992; Linn, 1989; Raju, 1992; Willson, 1989). Aunque no se ha estudiado de manera extensa la validez predictiva de la última ITBS, la evidencia de ediciones anteriores es muy alentadora. Por ejemplo, las calificaciones de la ITBS se correlacionan de manera moderada con las calificaciones en bachillerato (las *r* se encuentran alrededor de .60). La ITBS no es un instrumento perfecto, pero representa lo mejor que pueden producir los métodos actuales de desarrollo de pruebas.

PRUEBA METROPOLITAN DE RENDIMIENTO

La *Metropolitan Achievement Test*, MAT (Prueba Metropolitan de Rendimiento) data de 1930, cuando este instrumento se diseñó para satisfacer las necesidades de evaluación curricular de la ciudad de Nueva York. El propósito expreso de la MAT consiste en "medir el rendimiento de los alumnos en las principales áreas de habilidad y contenido del currículo escolar". Las normas de la MAT se obtuvieron de manera simultánea con las de la *Otis-Lennon School Ability Test*, OLSAT (Prueba de Capacidad Escolar Otis-Lennon).

La MAT es una batería multinivel diseñada para los grados escolares K a bachillerato. El instrumento consta de dos componentes que pueden utilizarse separados o combinados; ambos incluyen la Batería de Reconocimiento y las Baterías Diagnósticas. La primera proporciona información referida a normas acerca de habilidades académicas tradicionales. Por ejemplo, las subpruebas en el nivel de bachillerato (grados 10 a 12) incluyen las siguientes:

Lectura
Matemáticas
Lenguaje
Ciencias naturales
Ciencias sociales
Habilidades de investigación

La calificación de Habilidades de investigación se basa en reactivos relacionados con otras subpruebas. También se proporcionan calificaciones totales por batería.

Las Baterías diagnósticas consisten en tres pruebas diseñadas como herramientas para la planificación educativa; éstas encierran un enfoque referido a criterio para medir habilidades fundamentales en lectura, matemáticas y lenguaje. Por ejemplo, la prueba diagnóstica de lectura para los grados K-2 incluye medidas de discriminación visual, reconocimiento de letras, discriminación auditiva, vocabulario visual, fonemas/grafemas, vocabulario dentro de un contexto y comprensión de lectura. La MAT posee una confiabilidad excepcional, pero los críticos han señalado que sería deseable mayor evidencia de la validez relacionada con el criterio (Nitko, 1989; Rogers, 1989).

PRUEBAS IOWA DE DESARROLLO EDUCATIVO

Las *Iowa Tests of Educational Development*, ITED (Pruebas Iowa de Desarrollo Educativo),

que tienen un amplio uso, se publicaron por primera vez en 1942, y después se revisaron y reestandarizaron en varias ocasiones. El propósito de la ITED es "evaluar las habilidades intelectuales importantes en la vida adulta y proporcionar la base para el aprendizaje continuo". A diferencia de muchas otras pruebas de rendimiento que enfatizan habilidades vinculadas con metas curriculares específicas, la intención de la ITED es medir las metas fundamentales o habilidades generalizadas de educación independientes del currículo escolar; por tal razón, sus reactivos subrayan habilidades de orden superior. En vez de someter a prueba trozos aislados de conocimiento, las preguntas de la ITED presentan problemas que requieren la síntesis de conocimiento o una solución de pasos múltiples (figura 8–6).

La ITED es un instrumento diseñado para estudiantes de bachillerato. La prueba produce siete calificaciones básicas además de dos calificaciones derivadas:

Corrección e idoneidad de la expresión.
Vocabulario.
Capacidad de pensamiento cuantitativo.
Análisis de materiales en ciencias sociales.
Análisis de materiales en ciencias naturales.
Capacidad para interpretar materiales literarios.
Usos de fuentes de información.
Puntuación mixta.
Total de lectura.

La puntuación mixta se basa en todas las subpruebas, mientras el Total de lectura se obtiene de las subpruebas que requieren el análisis de material escrito. La ITED se afianza en ediciones anteriores, de modo que sin importar la forma, nivel o edición, una calificación dada representa el mismo nivel de excelencia. El pro-

Ciencias sociales

Anuncio

> Cuatro de cada cinco médicos entrevistados prefirieron
> JABÓN BÁLSAMO
> ¡Las pruebas demuestran que Jabón Bálsamo mejora los problemas de la piel más rápido que cualquier otro producto!

Con base en este anuncio, ¿cuál de las siguientes conclusiones, si es que alguna, es válida?

A. Se ha demostrado científicamente que la manera más rápida de librarse de cualquier problema de la piel es utilizar Jabón Bálsamo.

B. De las cinco marcas principales de jabones para la piel, sólo uno es mejor que Jabón Bálsamo desde el punto de vista médico.

C. De todos los médicos que recomendaron productos para el cuidado de la piel, cuatro de cada cinco recomendaron Jabón Bálsamo.

D. Ninguna de las conclusiones es válida.

Ciencias naturales

Poco después de la picadura de un mosquito, una persona enfermó de fiebre amarilla, ¿Cuál conclusión, si es que alguna, se puede justificar solamente a partir de estas observaciones?

A. Existe evidencia insuficiente para derivar cualesquiera de las conclusiones siguientes.

B. Los mosquitos son causa directa de la fiebre amarilla.

C. El mosquito introdujo un microorganismo en el torrente sanguíneo de la persona.

D. El mosquito portaba un organismo que causó la fiebre amarilla.

Figura 8–6. Reactivos representativos de las Pruebas Iowa de Desarrollo Educativo.
Fuente: Reimpreso de *Teacher, Administrator, and Counselor Manual: Iowa Tests of Educational Development, Forms X-8 and Y-8* (1988). Chicago, Il Riverside Publishing Co. Copyright © 1988. Reproducido con autorización de la Riverside Publishing Company.

pósito de vincular una prueba con ediciones anteriores consiste en facilitar las comparaciones longitudinales de crecimiento dentro de los individuos y simplificar la investigación a largo plazo de las tendencias históricas y educativas. En 1992, la ITED se reestandarizó con una considerable muestra nacional de estudiantes de bachillerato.

PRUEBAS VOCACIONALES Y DE APROVECHAMIENTO

Los *Tests of Achievement and Proficiency*, TAP (Pruebas Vocacionales y de aprovechamiento), son un instrumento diseñado para proporcionar una valoración amplia del progreso del alumno hacia las metas académicas tradicionales en bachillerato. La TAP es el segundo componente del *Riverside Basic Skills Assessment Program* (Programa Riverside de Evaluación de Habilidades Básicas); el primer componente lo forman las ITBS, que se emplean en los grados K a 9 (y que se analizaron anteriormente). Como la ITBS, las normas de la TAP se establecieron de modo simultáneo con las normas de la CogAT, una prueba de capacidad que mide las capacidades de razonamiento verbal, cuantitativo y no verbal. Las subpruebas de la TAP miden rendimiento en comprensión de lectura, matemáticas, expresión escrita, empleo de fuentes de información, ciencias sociales y ciencias naturales. También se proporciona una puntuación total o mixta.

La TAP también produce una calificación de aprovechamiento aplicado que evalúa la capacidad del alumno para manejar situaciones reales. Esta calificación refleja la competencia del alumno en la aplicación de habilidades matemáticas y de comunicación para resolver problemas de la vida cotidiana. Los reactivos enfatizan comunicación de ideas por escrito, solución matemática de problemas, uso de materiales de referencia e interpretación de materiales en tablas y gráficas. Las normas de la TAP se establecieron junto con las de la ITED y la CogAT, y se reestandarizó en 1992.

PRUEBAS DE DESARROLLO EDUCATIVO GENERAL

Otra batería de prueba muy utilizada la constituyen los *Test of General Educational Development*, GED (Pruebas de Desarrollo Educativo General), desarrolladas por el Consejo Estadounidense de Educación y que se aplican a nivel nacional en EUA para obtener el certificado de equivalencia de estudios secundarios (GED Testing Service, 1991). La GED consiste en exámenes de opción múltiple en cinco áreas educativas: habilidades de redacción, ciencias sociales, ciencias naturales, interpretación de literatura y artes y matemáticas. La sección de habilidades de redacción también contiene una pregunta de ensayo que las personas evaluadas deben responder por escrito. Dos lectores entrenados califican de manera independiente la pregunta de ensayo según un método de calificación holística de seis puntos. Los lectores deben hacer un juicio sobre el ensayo con base en su eficacia general en comparación con la de otros ensayos.

La GED se presenta en 12 formas alternas. Las confiabilidades de consistencia interna para las subescalas se encuentran, en su mayor parte, por encima de .90, con sólo algunas que llegan como mínimo a .87. La última edición de la GED no presenta confiabilidades de formas alternas, pero los resultados de versiones anteriores se encontraban en gran medida desde la mitad hasta la parte superior del rango de .80, con excepción de habilidades de lectura (que ahora se denomina interpretación de literatura y artes), para la cual la confiabilidad llegó a un nivel mínimo de .76. La confiabilidad intercalificadores para las muestras de redacción se localizó entre .6 y .7; estos datos indican que es apropiado aplicar un criterio liberal para la aprobación en esta subprueba, de modo que se reduzcan los errores de decisión. Con respecto a la validez, la GED tiene gran correlación con la prueba de lectura al momento de la graduación que se emplea en Nueva York ($r = .77$). Lo que es más, las normas para aprobar la GED son más rigurosas que aquéllas empleadas en la mayor parte de las escuelas de educación media.

En 1985, entre 27 y 33% de los estudiantes que se graduaron de preparatoria a nivel nacional hubieran sido incapaces de aprobar las Pruebas GED (Whitney, Malizio y Patience, 1985).

La GED enfatiza conceptos amplios en vez de hechos y detalles específicos. En general, su propósito es permitir que los adultos que no terminaron el bachillerato prueben que han obtenido un nivel equivalente de conocimiento a partir de sus experiencias o de estudios independientes. Los empleadores consideran a la GED como un equivalente (si no es que superior) a obtener un diploma de educación media superior. El desempeño exitoso de esta prueba permite que los individuos hagan solicitudes a universidades, busquen empleos y soliciten promociones que requieren un diploma de educación media superior. Rogers (1992) y Trevisan (1992) proporcionan reseñas inusualmente completas de la GED.

OTRAS PRUEBAS ESTANDARIZADAS DE RENDIMIENTO PARA GRUPOS

Además de las baterías que se listan antes, algunas pruebas estandarizadas de rendimiento que se emplean de manera grupal merecen un comentario breve. Dado que éstas se asemejan en gran medida a los instrumentos que se analizaron antes, se proporciona sólo un breve listado. Las *Sequential Tests of Educational Progress*, STEP-III (Pruebas Secuenciales de Progreso Educativo), se organizan en dos baterías, una que se utiliza con los grados K-3 y la otra que se emplea para los grados 3-12. La batería básica STEP-III evalúa las siguientes habilidades educativas: lectura, escritura, vocabulario, cálculo matemático y conceptos matemáticos. Pruebas adicionales miden logros en ciencias sociales, ciencias naturales, habilidades de estudio y comprensión oral (habilidades auditivas). La STEP-III es una prueba complementaria de las *School and College Ability Tests*, SCAT-III (Pruebas de Capacidad Escolar y Universitaria).

La muy utilizada *Stanford Achievement Series* (Serie de Pruebas de Ejecución Stanford) es uno de los programas de prueba más antiguos y prestigiosos en EUA. La serie incluye tres baterías relacionadas que cubren los grados K-13: la *Stanford Early School Achievement Test*, SESAT (Prueba Stanford de Rendimiento Escolar Temprano), para niños en jardín de niños y primer grado, la *Stanford Achievement Test*, SAchT (Prueba de Ejecución Stanford), para niños del primer grado a bachillerato, y la *Stanford Test of Academic Skills*, TASK (Prueba Stanford de Habilidades Académicas), para bachillerato y primer año de universidad. Los críticos son cautos en cuanto a la SESAT debido a que el valor de la prueba se fundamenta solamente en su validez de contenido. Poco se conoce acerca de su confiabilidad *test-retest*, su validez relacionada con el criterio y su validez de constructo (Ackerman, 1992; Carpenter, 1992). La SAchT se tiene en gran estima debido a su excelente cobertura referida a normas de un currículo educativo consensado, equilibrado y representativo a nivel nacional (Brown, 1992; Stoker, 1992). La TASK tiene excelentes características psicométricas pero, en su intento por abarcar el rendimiento en bachillerato y universidad, esta prueba emprende una tarea difícil. Después de todo, existe un modesto acuerdo en cuanto a las intenciones curriculares de la primaria, secundaria y preparatoria, pero ¿cuáles son las metas educativas del primer año de universidad?

PRUEBAS DE RENDIMIENTO CON PROPÓSITOS ESPECÍFICOS

Las pruebas de rendimiento pueden utilizarse para muchos propósitos aplicados importantes, incluyendo la estimación del conocimiento en campos avanzados y la evaluación de la competencia profesional. En esta última sección se examinarán dos pruebas de rendimiento con propósitos específicos. El *College-Level Examination Program* CLEP (Programa de Exámenes Universitarios), es un programa ampliamente utilizado a través del cual los estudiantes pueden demostrar su rendimiento a nivel universitario y recibir crédito o exención en ciertos cursos

universitarios. El *National Teacher Examination*, NTE (Examen Nacional para Maestros), es una prueba polémica que requieren muchos estados para extender la certificación a los maestros.

PROGRAMA DE EXÁMENES UNIVERSITARIOS

El CLEP es uno de dos programas nacionales de pruebas a través del cual los estudiantes pueden obtener créditos universitarios mediante un examen, sin inscribirse en los cursos. El otro programa es el *ACT Proficiency Examination Program* (Programa ACT de Examen de Capacidades), el cual no se analizará aquí. La *College Board* (Junta Universitaria) aplica el CLEP con respaldo económico de la *Carnegie Corporation* de Nueva York. El propósito original del programa era apoyar a estudiantes no tradicionales como los veteranos que regresaban del frente y los adultos mayores que habían obtenido valiosas experiencias de aprendizaje fuera del salón de clases. No obstante, son principalmente los ambiciosos estudiantes de educación media superior quienes se registran ahora para realizar los exámenes del CLEP. Algunos estudiantes comienzan la universidad con casi un semestre completo de créditos obtenidos a través del CLEP y programas similares.

El CLEP consiste en dos categorías de exámenes. Cada uno de los cinco exámenes generales es una prueba del material que se imparte en los cursos básicos que toma la mayoría de los estudiantes como parte de los requisitos generales de educación. Las áreas evaluadas incluyen composición en inglés, humanidades, matemáticas universitarias, ciencias naturales y ciencias sociales/historia. Excepto por la prueba de composición en inglés, que incluye un ensayo de 45 minutos, las pruebas generales son exámenes de opción múltiple con duración de 90 minutos. Las calificaciones satisfactorias en cada prueba general proporcionan por lo común de 3 a 6 horas de crédito semestral.

Cada uno de los 30 Exámenes Temáticos es una prueba de opción múltiple con duración de 90 minutos que se aplica en dos secciones cronometradas por separado. Cuatro de estas pruebas también incluyen una sección de ensayo que es requisito para algunas universidades. Pruebas de muestra incluyen temas sobre literatura estadounidense, francesa, alemana, española, gobierno estadounidense, macroeconomía, introducción a la psicología, álgebra, biología, química, ciencias informáticas y contaduría. Las calificaciones satisfactorias en cada prueba temática proporcionan tres o más horas de crédito semestral.

Las calificaciones de los Exámenes Generales del CLEP se presentan en una escala de 200 a 800 puntos, con un promedio de 500 y desviación estándar de 100. Para los Exámenes Temáticos, las calificaciones van en una escala de 20 a 80, con un promedio de 50 y desviación estándar de 10. Los grupos de referencia para éstas fueron estudiantes voluntarios que terminaron los cursos en las áreas especificadas. Éstos se reclutaron de una selección no aleatoria, pero supuestamente representativa, de colegios y universidades de EUA. En general, las calificaciones tienen confiabilidad elevada, con coeficientes por mitades que se colocan principalmente en el rango del .90. La validez de los Exámenes Temáticos se ha evaluado al correlacionar las calificaciones del CLEP con aquéllas finales en los cursos relacionados. La mayor parte de los coeficientes de correlación se encuentra en los rangos de .40 y .50, lo cual sustenta la validez concurrente de estas pruebas. Se supone que la validez de los Exámenes Generales es buena, pero se le ha establecido de manera menos precisa.

El programa CLEP ha recibido comentarios satisfactorios de los revisores, pero también tiene un lado potencialmente negativo. En particular, algunos de los estudiantes podrían pasar la prueba y no tomar algunos cursos universitarios que hubieran resultado enriquecedores, inspiradores e incluso hubieran cambiado su vida. Por ejemplo, es posible tener conocimientos objetivos acerca de arte, música o arte dramático y, por tanto, aprobar una prueba del CLEP en una o más de estas áreas. No obstante, en la búsqueda ambiciosa por terminar con rapidez la universidad, los estudiantes podrían pasar por alto importantes experiencias para su desarrollo personal.

EXAMEN NACIONAL PARA MAESTROS

El *National Teacher Examination*, NTE (Examen Nacional para Maestros) es, de hecho, una serie de pruebas publicadas por el *Educational Testing Service* y se conoce de manera más formal como *Praxis Series* (Serie Práctica). De los 40 estados que requieren la evaluación como parte del proceso para otorgar licencias, 36 utilizan la Serie Práctica, lo cual explica por qué la prueba se conoce de modo informal como *National Teacher Examination* o NTE. La Serie Práctica se aplica a nivel nacional y se actualiza y mejora continuamente. Las tres categorías de evaluación corresponden a los principales hitos en el desarrollo de la profesión en el área de enseñanza:

- Práctica I (Evaluación de habilidades académicas): ingreso a un programa de entrenamiento como maestro.
- Práctica II (Evaluación de materias): graduación de una universidad e ingreso a la profesión.
- Práctica III (Evaluaciones del desempeño en el salón de clases): primer año de enseñanza.

La prueba inicial, Práctica I, se realiza al inicio de la carrera universitaria del estudiante para evaluar habilidades de lectura, redacción y matemáticas, esenciales para el éxito de cualquier maestro. Se requiere una calificación aprobatoria en esta prueba de opción múltiple antes de que el estudiante pueda continuar con su especialización en educación. Estas pruebas pueden responderse en el formato tradicional de lápiz y papel o como una prueba para computadora que se diseña para el desempeño continuo de cada candidato. Una ventaja de la evaluación por computadora es su disponibilidad en todo momento, mientras que la versión tradicional sólo se aplica seis veces por año. Práctica II evalúa el conocimiento que tiene un candidato sobre la materia que impartirá, así como su conocimiento sobre la enseñanza de la materia. Se encuentran disponibles más de 120 áreas de contenido (todas de opción múltiple). Práctica III es una valoración en clase que realizan examinadores locales entrenados, los cuales utilizan criterios estructurados validados a nivel nacional.

La confiabilidad de las pruebas Práctica I y Práctica II está más allá de cualquier duda. De manera similar, la validez de contenido de estas pruebas es sobresaliente, debido a que se les construyó y refinó de manera cuidadosa con ayuda de muchos expertos y consumidores de pruebas. Lo que es menos claro es la validez predictiva de la Serie Práctica, ya que existe poca información para demostrar que las buenas calificaciones o valoraciones prácticas predicen una buena capacidad para la enseñanza o viceversa. Por supuesto, parte de la dificultad en este caso es encontrar una definición y medida adecuadas de lo que es la "buena capacidad para la enseñanza". Es probable que el NTE satisfaga un propósito útil al requerir que los futuros maestros posean niveles mínimos de conocimiento en sus disciplinas, pero la prueba también hace surgir preguntas difíciles relativas a la manera en que la sociedad identifica a los maestros prometedores. ¿Es suficiente el conocimiento objetivo? ¿No debería insistirse también en que los maestros posean entusiasmo por su material y capacidad para inspirar a los niños? Estos son aspectos que no se captan fácilmente a través de pruebas objetivas.

FALSEAMIENTO: EL LADO OSCURO DE LAS PRUEBAS DE RENDIMIENTO

La perspectiva prevaleciente entre el público en general es que el falseamiento ocurre en pocas ocasiones o nunca en los programas de prueba aplicados a nivel nacional. Se tiende a pensar que los riesgos son demasiado elevados y que quienes falsean los resultados tienen oportunidades demasiado limitadas como para tener éxito. Por tanto, se confía en que el fraude en las pruebas debe ser poco común; por desgracia, es probable que este punto de vista sea ingenuo. Después de todo, un creciente número de personas debe aprobar un examen para ingresar a la universidad, obtener un empleo o lograr un

ascenso. Lo que es más, cada vez es más frecuente que se evalúe a los funcionarios escolares con base en las calificaciones promedio de prueba en su distrito. Precisamente debido a los fuertes intereses que existen, individuos poco escrupulosos intentarán engañar al sistema.

Considérese el caso de las calificaciones superiores de una prueba en una aclamada escuela primaria en Connecticut (Associated Press, 4 de mayo de 1996 y 15 de marzo de 1997). La reputación sobresaliente de la escuela se basó en las altas calificaciones de examen obtenidas con la ITBS, que se aplicó a los niños de primero, tercero y quinto grados. La escuela había ganado preseas del *Education Department* de EUA y se describió como una de las mejores escuelas primarias del país en una destacada revista. No obstante, por un descubrimiento casual, el personal del distrito escolar observó un gran número de borraduras en las pruebas de esta escuela y notificó al editor de la prueba. Al inspeccionarlas de manera más cuidadosa, los editores encontraron un número excesivamente elevado de borraduras —9%— lo cual era de 3 a 5 veces mayor que en dos escuelas vecinas. Aún más sospechoso resultó el hecho de que 89% eran cambios de la respuesta incorrecta a la correcta. Con base en una repetición de la prueba bajo estrecha supervisión, la editorial de la prueba encontró "de manera clara y concluyente" que había alteraciones. El director renunció bajo acusaciones de que él era el responsable de la manipulación.

En 1997 se descubrió en Louisiana un caso especialmente flagrante de falseamiento en pruebas nacionales. Este caso implicó la circulación en masa del examen del *Educational Testing Service* (ETS) que se aplicaba a los maestros que deseaban ser directores de una escuela. Como se informó en el *New York Times* (28 de septiembre de 1997), copias de los 145 reactivos de la prueba, junto con las respuestas correctas, habían circulado entre los maestros a lo largo de todo el sur de Louisiana, muy probablemente durante varios años. En un estado clasificado como el último o casi último en el índice educativo, parecía que muchas personas potencialmente poco calificadas habían cometido fraude para obtener la dirección de las escuelas. El ETS manejó este caso de manera discreta al pedirle a más de 200 maestros que volvieran a realizar la prueba para "confirmar" sus calificaciones iniciales. Por desgracia, el caso de Louisiana no es un suceso aislado. El artículo del *New York Times* incluye esta inquietante conclusión:

> En numerosos incidentes a lo largo del país, el ETS ha confrontado un caso tras otro de falseamiento, pero le ha ocultado esta información al público y no ha tomado medidas agresivas para garantizar la integridad de sus pruebas, según lo demuestran documentos internos y entrevistas con sus funcionarios actuales y pasados.

Entre los ejemplos citados, se supone que el ETS no supervisó el manejo de la prueba del gobierno federal para inmigrantes que desean obtener la ciudadanía, con el factible resultado de que los supervisores de la prueba aceptaron sobornos. Las pruebas de dominio del inglés para estudiantes extranjeros también fueron vulnerables al falseamiento. En 1994, el ETS canceló las calificaciones de 30 000 estudiantes de China después de descubrir una red delictiva que vendía los exámenes en el exterior. En otro caso, procuradores federales descubrieron una red dedicada al falseamiento de pruebas en la que participaban cientos de estudiantes que pagaban miles de dólares cada uno por las respuestas del GRE y exámenes similares. Este fraude implicó una conocida estafa con husos horarios en la que personas experimentadas en realizar pruebas pasaban los exámenes en la ciudad de Nueva York y después enviaban las respuestas a clientes que contestaban las pruebas en husos horarios posteriores.

Las prácticas deshonestas e inapropiadas de los funcionarios escolares tienen una participación en la reciente inflación de las calificaciones de las pruebas grupales de rendimiento que tienen normas nacionales. Por definición, en el caso de una prueba referida a normas, 50% de las personas evaluadas debe tener calificaciones por encima del percentil 50, y 50% por debajo de ese percentil. Si se utiliza la misma prueba con una muestra amplia de sistemas

escolares típicos y representativos, las calificaciones promedio de los sistemas escolares deberían dividirse de manera uniforme —cerca de la mitad por encima de la norma nacional del percentil 50 y la otra mitad por debajo de éste.

Según una reciente encuesta presentada en los medios noticiosos (Foster, 1990), casi todos los estados de EUA afirman que las calificaciones promedio de rendimiento para sus sistemas escolares superan el percentil 50. La imagen demasiado optimista acerca del aprovechamiento estudiantil que resulta de ello se denomina **Efecto del Lago Wobegon**, con referencia al mítico pueblo de Minnesota, descrito por el humorista Garrison Keillor, donde "todos los niños son superiores al promedio".

¿Cómo surge la inflación de las calificaciones en pruebas de rendimiento? Según Cannell (1988), la causa principal son los administradores educativos que están desesperados por demostrar la excelencia de sus sistemas escolares. Precisamente debido a que la sociedad estadounidense le da tanta importancia a los resultados de las pruebas de rendimiento, en apariencia algunos educadores ayudan a sus alumnos a falsear las pruebas estandarizadas. El supuesto falseamiento incluye los siguientes puntos:

- Maestros y directores entrenan a los alumnos en las respuestas para la prueba.
- Los examinadores conceden más tiempo del asignado para contestar las pruebas.
- Los administradores alteran las hojas de respuesta.
- Los maestros enseñan directamente con referencia a los reactivos específicos de prueba.
- Los maestros hacen copias de las pruebas para proporcionarlas a sus alumnos.

Cannell señala que más de 300 maestros y administradores escolares respondieron a su anuncio en una publicación periódica del ramo y admitieron que ellos o sus colegas habían alterado las pruebas o ayudado a los estudiantes de maneras inapropiadas. Estas faltas constituyen una crisis silenciosa que continúa sin disminuir. Otra consecuencia del Efecto del Lago Wobegon es la probabilidad de que los editores de pruebas y los revisores federales aumenten sus esfuerzos por supervisar la seguridad de las pruebas. En resumen, la importancia que la sociedad le adjudica a las calificaciones en las pruebas de rendimiento ha provocado diversos efectos secundarios desagradables que socavan los fundamentos de los programas de pruebas grupales con normas nacionales.

Moore (1994) informa sobre un caso especial en relación con las evaluaciones educativas, a saber, las consecuencias a nivel de los distritos de las pruebas de rendimiento obligatorias por órdenes de un tribunal. Este autor hizo una encuesta con 79 maestros de tercero a quinto grado en un pueblo del medio oeste en el que los tribunales requerían una prueba estandarizada para determinar la eficacia de un esfuerzo para eliminar la segregación. La prueba en cuestión, ITBS, es una prueba grupal de rendimiento muy respetada que requiere el seguimiento estricto de instrucciones y límites de tiempo para obtener resultados válidos (véase antes). No obstante, los maestros le concedían poco valor al programa de pruebas y se quejaban de que sus beneficios no superaban el tiempo y costos implicados. Como consecuencia de su desprecio por este esfuerzo, la evaluación no estandarizada era prácticamente la regla más que la excepción. Los maestros realizaban diversas prácticas no estandarizadas, la mayor parte de las cuales tendían a inflar las calificaciones de la prueba. Las prácticas inapropiadas de examen incluían elogiar durante la prueba a los estudiantes que respondían de manera correcta una pregunta (67%), utilizar preguntas de la prueba del año anterior como práctica (44%), volver a anotar la clave en la hoja de respuestas de un alumno porque el estudiante simplemente había "confundido" la respuesta (26%), dar a los estudiantes tanto tiempo como necesitaran (24%), aplicar a los estudiantes reactivos que estaban directamente fuera de la prueba (24%) y dar indicios o claves durante la prueba (23%). En general, Moore (1994) señala que los maestros modificaron tanto sus esfuerzos al dar instrucciones como el currículo de materias antes de que sus estudiantes contestaran la prueba. Más de 90% de los maestros

añadieron al currículo lecciones relacionadas con la prueba, y más de 70% eliminaron temas, de modo que pudieran tener más tiempo para ocuparlo en habilidades relacionadas con la prueba. Con toda seguridad, el que estos cambios sean deseables está abierto a debate. Moore (1994) concluye:

> La evaluación estandarizada ha tenido un papel central en la educación durante muchos años. Lo que han demostrado recientemente los estudios sobre el impacto de los programas de evaluación es la creciente dependencia en las calificaciones de una prueba para la toma de decisiones y el potencial cada vez mayor de uso impropio de las mismas. Es necesario que aquellos que determinan los procedimientos en educación y política atiendan al importante vínculo entre instrucción y prueba y que garanticen que los maestros se integren a los esfuerzos de evaluación en lugar de aislarse de ellos.

En resumen, lo que demuestra este estudio es que las pruebas educativas obligatorias pueden tener la consecuencia no anticipada de contaminar la validez de una prueba meritoria —en especial cuando los principales interesados no tienen opinión dentro del proceso.

En nuestro caso no es posible analizar todos los efectos colaterales no intencionales de las pruebas de rendimiento educativo, debido a que las posibilidades son casi infinitas. Por ejemplo, ¿qué sucede con los efectos modificadores de los programas de pruebas de rendimiento sobre los currículos escolares? Como se ha visto en el estudio de Moore (1994), los maestros cambian sus prácticas de clase para ayudar a los alumnos a obtener buenas calificaciones en las pruebas; no obstante, al enseñar con el propósito de pasar la prueba, los educadores pueden enfatizar el conocimiento objetivo en vez de impartir una capacitación general para pensar con claridad y resolver problemas. En conclusión, parece ser que el énfasis excesivo en las pruebas de rendimiento con normas nacionales para la selección y evaluación promueve conductas inapropiadas, incluyendo fraude evidente y falseamiento de parte de alumnos y funcionarios escolares. ¿Qué tan extendido está el problema? Aunque se vive con la suposición optimista de que el fraude en los programas nacionales de prueba es poco común, la perturbadora verdad es que en realidad no se sabe con qué frecuencia ocurre.

RESUMEN

1. Las pruebas de rendimiento se emplean para medir el logro de las habilidades que se enseñan dentro de escuelas o programas de entrenamiento. Las pruebas grupales de rendimiento se utilizan principalmente en el salón de clases, mientras las individuales se emplean en entornos clínicos o educativos individuales.
2. La distinción entre capacidad, aptitud y rendimiento es confusa. De hecho, las correlaciones entre estos tres tipos de medidas son, con frecuencia, muy elevadas; no obstante, las aplicaciones típicas de estas pruebas son diferentes: las pruebas de aptitud se utilizan para predecir el desempeño futuro, mientras las de rendimiento estiman el funcionamiento actual.
3. Las funciones de las pruebas de rendimiento educativo incluyen la detección de una posible discapacidad para el aprendizaje, identificación de fortalezas y debilidades individuales, agrupamiento de estudiantes y valoración del éxito de programas educativos.
4. Los editores de pruebas han desarrollado varias baterías multinivel para satisfacer las necesidades que tienen los sistemas escolares de evaluar el rendimiento educativo. En una batería multinivel, los protocolos de prueba contienen secciones superpuestas y los estudiantes en diferentes niveles de grado inician y terminan con los materiales de la prueba en los lugares apropiados para su grado.

5. Las ITBS constituyen una batería multinivel de pruebas de rendimiento que cubre los grados K-9. Sus normas se obtuvieron al mismo tiempo que las de la CogAT, y la ITBS mide el rendimiento en áreas básicas como vocabulario, comprensión de lectura, ortografía y matemáticas.
6. La MAT consta de dos componentes que pueden utilizarse por separado o en combinación: Batería de Reconocimiento (evaluación referida a normas de habilidades académicas tradicionales) y Baterías Diagnósticas (enfoque referido a criterio con propósitos de planificación educativa).
7. A diferencia de muchas otras pruebas de rendimiento que enfatizan habilidades ligadas con metas curriculares específicas, la intención de las ITED es medir las metas fundamentales o habilidades generalizadas de la educación independientes del currículo de materias. Por tal razón, los reactivos empleados en la ITED destacan habilidades de pensamiento de orden superior.
8. Las TAP se diseñaron para proporcionar una evaluación amplia del progreso del alumno hacia las metas académicas tradicionales de bachillerato. La TAP es una prueba relacionada con la ITBS y sus normas se obtuvieron de manera simultánea con las de la CogAT, una prueba que mide capacidades de razonamiento verbal, cuantitativo y no verbal.
9. Otra batería de pruebas de rendimiento de empleo generalizado la constituyen las GED, desarrolladas por el *American Council on Education* y que se aplican a nivel nacional para obtener el certificado de equivalencia de bachillerato. La GED consiste en exámenes de opción múltiple en cinco áreas educativas, e incluye preguntas de ensayo para evaluar habilidades de redacción.
10. El CLEP permite que los estudiantes obtengan créditos universitarios a través de un examen sin inscribirse a los cursos. Existen cinco exámenes generales (p. ej., humanidades, matemáticas universitarias) y 30 exámenes específicos (p. ej., literatura estadounidense, introducción a la psicología, álgebra, biología).
11. El NTE, que publica el *Educational Testing Service*, se conoce de manera más formal como Serie Práctica. Esta valoración consiste en tres categorías (habilidades académicas, conocimiento de la materia y desempeño en el salón de clase). Las evaluaciones temáticas son un requisito previo en muchos estados para obtener la licencia.
12. Se desconoce la incidencia del falseamiento en pruebas de rendimiento de aplicación nacional. Sin embargo, en años recientes han surgido muchos informes que incluyen la alteración de hojas de respuesta por parte de funcionarios escolares, la circulación en masa de algunos exámenes para licencias y las prácticas inapropiadas de prueba de algunos maestros individuales (p. ej. conceder tiempo adicional para terminar la prueba).

TÉRMINOS Y CONCEPTOS CLAVE

Pruebas grupales de rendimiento p. 348
Pruebas individuales de rendimiento p. 348
Efecto del Lago Wobegon p. 359

■

CAPÍTULO 9

Evaluación neuropsicológica y geriátrica

TEMA 9A
Compendio de neuropsicología

Anatomía del cerebro
Ejemplo de caso 9–1. Disfunción cerebral y resultados de prueba
Metencéfalo
Mesencéfalo
Diencéfalo
Lóbulo límbico
Ganglios basales
Cuerpo calloso
Corteza cerebral
Funciones de los lóbulos cerebrales
Lateralización cerebral de la función
Funciones especializadas del hemisferio derecho
Pruebas clínicas y técnicas de imagenología cerebral
Aplicaciones de los hallazgos de pruebas neuropsicológicas
Resumen
Términos y conceptos clave

La evaluación neuropsicológica es la aplicación de pruebas especializadas con propósitos de diagnóstico y tratamiento de individuos en los que se sospecha o conoce disfunción cerebral. Las pruebas neuropsicológicas son específicas debido a su vínculo demostrado con las funciones cerebrales (Lezak, 1995). En términos específicos, cualquier prueba o técnica que sea sumamente sensible a los efectos del deterioro cerebral —y en especial una prueba o técnica que permita hacer inferencias sobre el lugar, tipo y grado del deterioro— tendría la categoría de procedimiento neuropsicológico. El propósito de este capítulo es introducir al lector a las pruebas, conceptos y métodos de este tipo. Además, se incluye un énfasis secun-

dario en la evaluación de los trastornos relacionados con la edad, como la enfermedad de Alzheimer, demencia con infartos múltiples (accidente cerebrovascular) y depresión de la vejez, debido a que estos problemas pueden comprenderse mejor dentro de un contexto neuropsicológico. En el Tema 9A, Compendio de neuropsicología, se proporciona una revisión condensada de las funciones cerebrales humanas. También se presta especial consideración a los trastornos del envejecimiento y se reseñan los enfoques médicos a la imagenología del cerebro. El propósito central de este tema es preparar al lector para proseguir con el análisis de los procedimientos neuropsicológicos y geriátricos de evaluación. En el Tema 9B, Evaluación neuropsicológica y geriátrica, se analiza la naturaleza y empleo de las principales pruebas neuropsicológicas y procedimientos de evaluación en geriatría.

La **neuropsicología** es el estudio de la relación entre función cerebral y conducta (Cytowic, 1996; Kolb y Whishaw, 1990); no floreció sino hasta la segunda Guerra Mundial, cuando los psicólogos y neurólogos se enfrentaron con miles de soldados que tenían lesiones cerebrales y que requerían de evaluación y rehabilitación (Aita, Armitage, Reitan y Rabinowitz, 1947; Goldstein, 1944). Era una circunstancia sombría para la innovación, pero la carnicería ocurrida en la guerra obligó a los psicólogos a desarrollar nuevas herramientas para comprender los efectos del daño cerebral. Así, mientras la humanidad devastaba a su propia especie durante esta guerra, varios especialistas forjaron una nueva especialidad, la neuropsicología clínica, para ayudar a manejar las consecuencias. El objetivo inicial y fundamental de esta novel disciplina consistió en determinar las relaciones entre la lesión cerebral y el desempeño en pruebas psicológicas (Ejemplo de caso 9–1).

El punto central de la neuropsicología es la promoción de una ciencia de la conducta humana fundamentada en la función cerebral humana (Kolb y Whishaw, 1990). Los estudiantes de la evaluación neuropsicológica deben estar familiarizados con los aspectos esenciales de la función cerebral, si es que han de apreciar el papel de las pruebas en el diagnóstico de los padecimientos relacionados con el deterioro cerebral. En este tema, se presenta un compendio de términos y conceptos neurológicos, incluyendo una revisión de la anatomía y función cerebrales, con un análisis de los enfoques médicos a la imagenología del cerebro. También se toca de manera breve el tema de los principales tipos de neuropatología que podrían alterar el funcionamiento normal del cerebro, con especial énfasis en los ancianos.

ANATOMÍA DEL CEREBRO

Por convención, el sistema nervioso se divide en Sistema Nervioso Central (que consiste en cerebro y médula espinal) y Sistema Nervioso Periférico (que incluye los pares craneales y la red de nervios que emana de la médula espinal). La evaluación neuropsicológica intenta determinar la relación entre cerebro y conducta, de modo que aquél será el principal punto de atención de este análisis.

El cerebro pesa cerca de 1.300 kg. Se forma principalmente de dos componentes, neuronas y células gliales. Durante decenios se ha creído que las neuronas no se reproducen; no obstante, investigaciones recientes sugieren que los humanos pueden tener una capacidad limitada para reproducir neuronas, en especial en las áreas del cerebro que son importantes para el aprendizaje y la memoria (*Associated Press*, 30 de octubre de 1998). Las más numerosas células gliales, que tienen la capacidad de reproducción prolífica, proporcionan diversas clases de apoyo estructural a las neuronas.

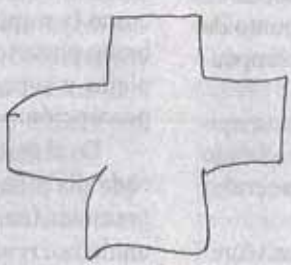
Los 10^{11} o 100 mil millones de neuronas en el cerebro se distribuyen en redes complejas que en gran medida han desafiado a la comprensión. En parte, la cualidad inescrutable del cerebro se deriva de su complejidad computacional. Las neuronas se comunican al enviarse impulsos electroquímicos de todo o nada, unas a otras. Cada una podría enviar transmisiones a miles, y quizá decenas de miles, de otras

EJEMPLO DE CASO 9–1
DISFUNCIÓN CEREBRAL Y RESULTADOS DE PRUEBA

Se pidió a un neuropsicólogo que evaluara a un estudiante de los primeros niveles de universidad, cuyas bajas calificaciones eran fuente de desconcierto para el asesor, los padres e incluso el psiquiatra de este alumno, quien había tenido un funcionamiento adecuado durante dos años en la universidad, obteniendo un promedio de 9 en un currículo variado de materias. No obstante, ahora estaba a punto de reprobar las materias previas a la carrera de medicina, las cuales incluían cursos sobre química orgánica y embriología. El alumno se quejaba de que el material en sus cursos de ciencias naturales estaban simplemente más allá de su entendimiento. Sin embargo, obtenía calificaciones de 8 en historia y literatura. Con base en el deterioro de sus calificaciones, el psiquiatra sospechó que el estudiante podría sufrir de un daño cerebral leve. Se solicitó una valoración neurológica.

Aquí no se revisará toda la batería de pruebas, pero ciertos resultados servirán para introducir el vínculo entre lo obtenido por una prueba y la función cerebral. Quizás el descubrimiento más importante fue que el alumno no pudo copiar adecuadamente el patrón geométrico sencillo de una cruz (véase ilustración). Otro dato sugerente era una enorme diferencia en control motor fino entre ambos lados de su cuerpo. Con un aparato estandarizado de golpeteo con los dedos, el neuropsicólogo estableció que el estudiante podía golpetear 56 veces en 10 segundos con su dedo índice derecho, pero sólo lograba hacerlo 42 veces con su dedo índice izquierdo (Morrison, Gregory y Paul, 1979). Finalmente, aunque su CI total era de 122, el estudiante tuvo un desempeño deficiente en la *Category Test* (Prueba de Categorías), una medida del razonamiento abstracto y de formación de conceptos que discrimina de manera adecuada entre personas con daño cerebral y neurológicamente intactas (Reitan y Wolfson, 1993).

Dibujo de la cruz griega:



Golpeteo con los dedos: Derecho 56 Izquierdo 42

CI Total en el WAIS-R 122 CI Verbal 133 CI de Ejecución 105

Prueba de categorías: 56 errores/208

¿Este estudiante tenía daño cerebral? De ser así, ¿cuál era el sitio más probable en el que había deterioro? ¿Cuáles serían las implicaciones de esta aparente disfunción? Se regresará al caso del estudiante reprobado de los cursos previos a la carrera de medicina al final de este tema. Hasta ahora, el lector poseerá el suficiente conocimiento acerca de las relaciones cerebro-conducta para comprender un análisis neuropsicológico de los resultados de prueba que se mostraron antes.

neuronas en sitios cercanos y distantes llamados sinapsis. Las comunicaciones químicas entre sinapsis pueden ocurrir hasta mil veces por segundo. Incluso si se utiliza un estimado conservador de 1 000 sinapsis por neurona, en teoría, el número de transmisiones neuronales que pueden ocurrir en sólo un segundo es un sorprendente 10^{17} o 100 000 000 000 000 000 (cien mil billones). Por ello, no es sorpresa que neurocientíficos serios como Sir John Eccles (quien recibió el Premio Nobel por su trabajo en neuropsicología) recurran a la hipérbole para describir al cerebro como "sin reservas, la materia orgánica más altamente organizada y con mayor complejidad en el universo" (Eccles, 1973).

A pesar de las dificultades para comprender al cerebro, los neurocientíficos han formulado algunas teorías rudimentarias y parciales a partir de las siguientes fuentes:

1. Estudios conductuales sobre pacientes con lesiones cerebrales separadas.
2. Observaciones clínicas de los efectos de la estimulación eléctrica en sitios seleccionados del cerebro en pacientes epilépticos en estado de conciencia y que están a punto de atravesar por escisiones cerebrales terapéuticas.
3. Estudios de laboratorio con pacientes epilépticos a quienes se les han desconectado quirúrgicamente los hemisferios cerebrales para controlar las crisis.
4. Estudios de investigación con nuevas técnicas de imagenología del cerebro que proporcionan un análisis en tiempo real de la actividad cerebral continua.

En lo que resta de este tema, se proporciona al lector un compendio de las relaciones cerebro-conducta que se han identificado a partir de estas fuentes.

ORGANIZACIÓN FUNCIONAL DEL CEREBRO

La organización funcional del cerebro humano es difícil de comprender debido a que las estructuras importantes están entrelazadas y plegadas unas sobre otras. Además, contiene un complicado sistema de cavernas llenas de líquido que se denominan **ventrículos**. Aunque comienzan en el centro del cerebro, éstos tienen un sistema de canales que en última instancia se extiende hacia fuera y envuelve a todo el cerebro, y también va hacia abajo a la médula espinal para actuar como amortiguador para las estructuras extremadamente delicadas del Sistema Nervioso Central.

Aunque la disposición espacial de las estructuras cerebrales es muy compleja (figura 9–1) desde la perspectiva de las interconexiones neuronales, las divisiones del cerebro son sucesiones lineales de una a otra: metencéfalo, mesencéfalo y prosencéfalo. Estructuralmente, los centros cerebrales inferiores (en el rombencéfalo) son los que tienen la organización más simple, mientras los centros cerebrales superiores (en el cerebro posterior) son grandes, elaborados y anatómicamente completos. La organización funcional sigue también este patrón: los centros cerebrales anteriores median las funciones primitivas más simples, como la respiración, mientras los centros cerebrales posteriores gobiernan funciones más complejas y superiores, como el pensamiento y la percepción.

En el cerebro completamente desarrollado, cada división consta de varias estructuras importantes (cuadro 9–1). En lo que resta de esta unidad se revisa la función y disfunción de estos componentes estructurales principales del cerebro humano.

METENCÉFALO

La parte más inferior del cerebro es el **metencéfalo** y su sección más baja es el bulbo raquídeo. Todas las fibras nerviosas que provienen de las partes superiores del cerebro atraviesan la **médula ósea** en camino a la médula espinal. El entrecruzamiento de estas fibras explica por qué las dos mitades del cerebro terminan controlando lados opuestos del cuerpo. La médula

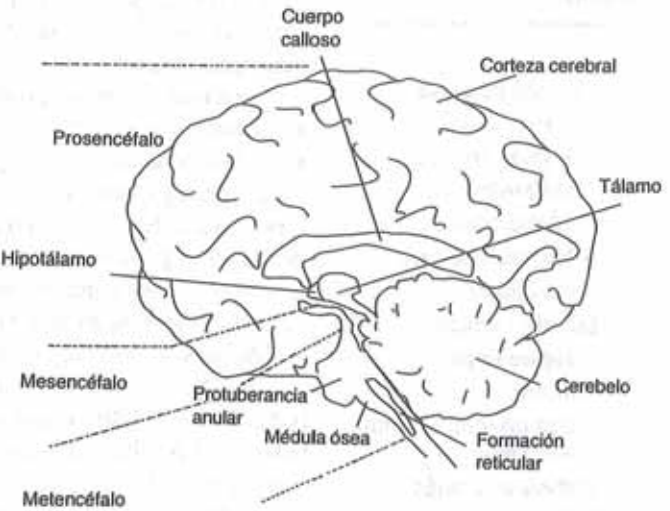


Figura 9–1. Perspectiva media del hemisferio cerebral derecho que muestra importantes estructuras en el prosencéfalo, mesencéfalo y metencéfalo.

ósea media varias funciones corporales esenciales: respiración, deglución, vómito, presión sanguínea y, en parte, frecuencia cardiaca (Kandel, Schwartz y Jessell, 1995). También se gobiernan aquí ciertos aspectos del habla y el canto, aunque sitios cerebrales superiores también participan en estas funciones.

Por lo general, el daño importante a la médula ósea resulta mortal. En casos aislados, una pequeña embolia en la médula ósea causa uno o más de los siguientes síntomas: parálisis del lado opuesto, pérdida parcial de la sensación al dolor y temperatura, torpeza, mareo, parálisis del mismo lado y atrofia de la lengua, y pérdida parcial del reflejo nauseoso. El virus de la polio —tan predominante en el decenio de 1950, pero que no se ha controlado totalmente— puede atacar la médula y provocar la suspensión del control neuronal para la respiración, lo cual hace que se requiera de un respirador mecánico.

La **formación reticular**, una red de cuerpos y fibras neuronales ascendentes y descendentes, comienza en la médula espinal y se extiende a través de la médula ósea en camino hasta el tálamo. Núcleos específicos dentro de la formación reticular se proyectan a amplias áreas del cerebro y, por tanto, ayudan a mediar reflejos posturales complejos y tono muscular. Con base en los estudios de Moruzzi y Magoun (1949) que demuestran que los tractos nerviosos ascendentes dentro de la formación reticular gobiernan la excitación general o estado de conciencia, partes de esta estructura también se conocen como sistema reticular activador. El daño a este sistema da lugar a la disminución global de la conciencia que va desde adormecimiento crónico hasta estupor y coma (Carpenter, 1991).

La protuberancia anular y el cerebelo son las estructuras más altas en el metencéfalo. Juntos ayudan a coordinar tono muscular, postura y movimientos oculares y manuales. Las lesiones a esta protuberancia pueden impedir que el individuo realice movimientos laterales coordinados con los ojos. Por tal razón, es común que los neurólogos pidan a sus pacientes que demuestren movimientos oculares de izquierda a derecha y de arriba hacia abajo.

Aunque muchos sitios cerebrales participan en el control motor, el **cerebelo** desempeña

Cuadro 9–1. Divisiones del cerebro humano

Prosencéfalo	
Telencéfalo (cerebro terminal)	Corteza cerebral
	Cuerpo calloso
	Ganglios basales
	Putamen
	Globo pálido
	Núcleo caudado
	Amígdala
	Lóbulo límbico
	Hipocampo
	Tabique
	Circunvolución cingulada
	Bulbos olfatorios
Diencéfalo (cerebro intermedio)	Tálamo
	Hipotálamo
	Glándula pineal
Mesencéfalo	
Mesencéfalo (cerebro medio)	Tectum
	Colículos superiores
	Colículos inferiores
	Tegmento
	Nervios craneales
Metencéfalo	
Metencéfalo (cerebro de intersección)	Cerebelo
	Protuberancia anular
	Formación reticular*
Mielencéfalo (cerebro espinal)	Médula ósea

* La formación reticular comienza en el telencéfalo y se extiende hacia arriba al diencéfalo.

una importante función. Éste recibe información sensorial de todas las partes del cuerpo y coordina los detalles de los movimientos automáticos hábiles. El daño al cerebelo puede producir una variedad de trastornos motores, dependiendo de los sitios específicos afectados (Bradshaw y Mattingley, 1995). Los músculos pueden volverse flácidos y cansarse con facilidad. Puede resultar difícil el golpeteo rápido, coordinado, del dedo índice. Por tanto, las medidas de velocidad de golpeteo con los dedos (Reitan y Wolfson, 1993) son un componente importante de las baterías de pruebas neuropsicológicas.

Los movimientos corporales pueden perder su coordinación cuando se daña el cerebelo, volviéndose espasmódicos y temblorosos. Incluso un simple ademán como tratar de alcanzar una taza puede terminar en arrojar inadvertidamente ésta y su contenido al otro lado de la habitación. La característica marcha irregular que se encuentra en muchos individuos alcohólicos crónicos —denominada ataxia— es una consecuencia de la degeneración del cerebelo (Ghez, 1991). Otro síntoma del daño al cerebelo es el temblor de intención, llamado así porque no está presente en reposo, sino que surge durante los movimientos voluntarios, intencionales, de las manos. El nistagmo también es común en la enfermedad del cerebelo. En este síntoma, los ojos parecen moverse de un lado a otro incluso cuando el individuo intenta mantener la mirada fija.

Junto con el centro vestibular en el oído interno, el cerebelo también ayuda a coordinar el reflejo vestibulocular (RVO), el cual actúa para mantener los ojos en un lugar fijo cuando se gira la cabeza. Sin el RVO, la visión sería increíblemente borrosa cuando se moviera la cabeza inclusive una fracción de un centímetro. En vez de ello, una pequeña área del cerebelo coordina la fijación rápida de los ojos para compensar los movimientos de la cabeza.

MESENCÉFALO

El **mesencéfalo** consiste en dos divisiones principales que cubren un acueducto lleno de líquido. El tectum o techo se encuentra por encima del acueducto, y el tegmento o piso se encuentra por debajo de éste. El tectum consta, principalmente, de dos conjuntos de núcleos bilateralmente simétricos, los colículos superiores y los colículos inferiores. Los primeros median los movimientos de cabeza y ojos que se utilizan para localizar y seguir estímulos

visuales. Los colículos inferiores proporcionan la misma función para los estímulos auditivos.

El tegmento es una estación de relevo para las fibras sensoriales y motoras y también contiene los núcleos de muchos de los nervios craneales (algunos de los cuales también parten del metencéfalo). Los 12 **pares craneales** son los principales haces neuronales cuyas funciones son bien específicas y pueden examinarse con facilidad. Algunos son exclusivamente sensoriales y envían información del mundo exterior al cerebro; algunos son sólo motores y sirven para ejecutar las órdenes de éste; cerca de un tercio de los pares craneales poseen tanto funciones sensoriales como motoras. Los neurólogos se refieren a ellos por números, que corresponden aproximadamente a la secuencia de arriba hacia abajo en que surgen a partir del cerebro (cuadro 9–2). El lector observará que muchos de los pares craneales median aspectos de visión y movimiento ocular, funciones sensoriales básicas y movimiento de mandíbula, lengua, rostro y cabeza. A través de los siglos, los neurólogos han diseñado una variedad de técnicas confrontativas simples para evaluar los pares craneales. Por peculiar que pueda parecer, pedirle al paciente que saque la lengua y la mueva a la izquierda, derecha, arriba y abajo, puede proporcionar información importante acerca del funcionamiento del par craneal hipogloso (doceavo par). De igual manera, varias pruebas sencillas de audición, equilibrio, movimiento ocular y demás se utilizan para llevar a cabo el examen de los pares craneales.

Cuadro 9–2. Pares de nervios craneales y sus funciones

1. Olfatorio	Sentido del olfato
2. Óptico	Visión
3. Oculomotor	Movimientos horizontal y vertical del ojo
4. Patético	Movimiento vertical del ojo
5. Trigémino	Sensación facial, movimiento de la mandíbula
6. Motor ocular externo	Movimiento horizontal del ojo
7. Facial	Movimiento facial y gusto
8. Auditivo/vestibular	Oído y equilibrio
9. Glosofaringeo	Gusto, deglución
10. Vago	Reflejos viscerales
11. Accesorio	Movimiento de la cabeza
12. Hipogloso	Movimiento de la lengua

Aunque los neuropsicólogos tienen algunas herramientas y procedimientos adecuados para la evaluación de la disfunción del metencéfalo y mesencéfalo, son los especialistas médicos, en particular los neurólogos, quienes por lo regular diagnostican y tratan las discapacidades relacionadas con estos sitios cerebrales. Las pruebas y procedimientos neuropsicológicos se concibieron principalmente para la evaluación de la función y disfunción en el **prosencéfalo**. Por tanto, se dedica la mayor parte de este trabajo a las estructuras de esta parte del cerebro. El prosencéfalo consta del diencéfalo, un conjunto de estructuras menores —que el puño— que se encuentran aproximadamente en el centro del cerebro, y el telencéfalo, un grupo más grande de estructuras cerebrales que están colocadas por encima del diencéfalo (cuadro 9–1).

DIENCÉFALO

Las estructuras del diencéfalo incluyen tálamo, hipotálamo y **glándula pineal**. Esta última es una estructura del tamaño de un guisante que se encuentra en el centro del cerebro, y se sabe que secreta la hormona melatonina con un ritmo biológico cíclico, pero su función exacta se desconoce (Crapo, 1985), por lo que los textos contemporáneos de neurociencias casi la ignoran. Desde la perspectiva de la neuropsicología, el tálamo y el hipotálamo son mucho más importantes.

El **tálamo** es una pequeña estructura bifurcada en la base del cerebro. No se puede

exagerar su importancia: es la estructura clave que proporciona información y datos sensoriales a la corteza cerebral acerca del movimiento continuo. De hecho, toda la información sensorial, excepto la olfatoria (el sentido del olfato), se envía primero al tálamo y después se proyecta a regiones específicas de la corteza cerebral a través de haces neuronales conocidos en conjunto como radiaciones talamocorticales. Una lesión amplia a un lado del tálamo causa alteración grave de todas las formas de sensibilidad en el lado opuesto del cuerpo. Las lesiones más pequeñas pueden producir que se eleve el umbral al dolor; además, los estímulos dolorosos pueden causar una respuesta exagerada. Con las lesiones al tálamo puede disminuir de manera notable la percepción de la postura y de los movimientos pasivos (Carpenter, 1991).

El tálamo es mucho más que una estación de relevo; también representa una función importante en la integración de los sistemas neuronales pues contribuye a la memoria, atención, habla y experiencia emocional (Kandel, Schwartz y Jessell, 1995). Carpenter (1991) afirma que el tálamo es la clave para comprender la operación de la corteza cerebral.

El **hipotálamo** es una estructura engañosamente pequeña que se encuentra justo por debajo y al frente del tálamo. Aunque constituye sólo cerca de 0.3% del peso del cerebro, participa en numerosos aspectos de la conducta motivada y de la regulación del cuerpo: alimentación, conducta sexual, sueño, regulación de temperatura, conducta emocional y movimiento. Bien estudiadas en animales inferiores, las funciones del hipotálamo se conocen menos en los humanos (Kolb y Whishaw, 1990). Se sabe que ejerce dominio sobre la glándula hipofisaria, con lo cual modula una gran variedad de funciones endocrinas. La causa más común de lesión al hipotálamo es un traumatismo grave en la cabeza. Las lesiones hipotalámicas conducen con frecuencia a alteraciones de la función de la hipófisis, incluyendo ingestión excesiva o deficiente de alimento o agua y problemas con la regulación de temperatura y presión sanguínea (Kupfermann, 1991a).

La formación más externa del prosencéfalo es el telencéfalo, una estructura mucho mayor que el diencéfalo. Las pruebas neuropsicológicas son particularmente adecuadas para detectar la disfunción en el telencéfalo. A continuación se presentan análisis independientes de las siguientes estructuras telencefálicas: lóbulo límbico, ganglios basales, cuerpo calloso y corteza cerebral. Por favor, véase que es inevitable cierto grado de simplificación excesiva cuando se estudian las funciones cerebrales. Se recuerda al lector que el cerebro tiene abundantes interconexiones con sistemas funcionales complejos y mal comprendidos. Los sitios cerebrales individuales pocas veces actúan en aislamiento.

LÓBULO LÍMBICO

El **lóbulo límbico** incluye al hipocampo, tabique y circunvolución cingulada. Estas estructuras revisten la pared interna de la corteza cerebral. El lóbulo límbico suele conocerse como sistema límbico, lo cual con toda probabilidad es una denominación incorrecta basada en especulaciones anteriores que ahora se saben incompletas. En particular, Papez (1937) propuso que las estructuras del lóbulo límbico constituyen un circuito neuronal continuo, responsable de la emoción y control de la actividad visceral —de aquí la referencia a sistema límbico. Esta perspectiva ha resultado demasiado simplista, pero la atractiva referencia a un sistema funcional aún continúa. Ahora parece claro que las estructuras independientes del lóbulo límbico pueden tener muchas funciones diferentes, incluyendo una actividad primordial en la regulación de la emoción (Kupfermann, 1991a).

Se sabe que una estructura límbica, el **hipocampo**, representa un papel especial en la memoria. Los humanos poseen dos hipocampos (uno en cada lóbulo temporal) que son parte de un circuito complejo y mal definido de memoria que consolida las experiencias nuevas en recuerdos a largo plazo (Mishkin y Appenzeller, 1987). El daño a un hipocampo puede producir deficiencias leves en la memoria; no obstante, las personas que han sufrido daño en ambos

hipocampos o que se los han retirado quirúrgicamente, experimentan incapacidad para recordar cualquier cosa nueva por más de pocos segundos (Milner, Corkin y Teuber, 1968). Estos desafortunados individuos conservan los recuerdos a largo plazo adquiridos antes de la destrucción bilateral de los hipocampos, pero no pueden convertir los recuerdos nuevos a corto plazo en memorias a largo plazo; dichas personas podrían leer la misma revista una y otra vez y saludar al doctor cada día como si fuera un desconocido. La inteligencia general se afecta poco; sin embargo, debido a su memoria deficiente, estos pacientes requieren de manera invariable institucionalización o un alto grado de supervisión.

La destrucción unilateral del hipocampo causa alteraciones en la memoria selectiva que se correlacionan con especializaciones hemisféricas bien conocidas (funciones verbales en el hemisferio izquierdo, funciones espaciales en el hemisferio derecho, las cuales se analizarán posteriormente). La pérdida del hipocampo izquierdo altera la memoria verbal, mientras la destrucción del hipocampo derecho afecta la memoria pictórica y auditiva, en particular con patrones visuales y auditivos que no pueden clasificarse con facilidad (Smith y Milner, 1981). En contraste con las perturbaciones de memoria producidas por el daño bilateral al hipocampo que son profundas y de larga duración, los trastornos de memoria causados por el daño unilateral tienden a ser transitorios y sutiles. Se requieren pruebas neuropsicológicas complejas para su detección (Rausch, 1985).

GANGLIOS BASALES

Los **ganglios basales** consisten en un conjunto de núcleos en el prosencéfalo que se conectan con la corteza cerebral arriba y el tálamo abajo; tradicionalmente se ha considerado que son parte del sistema motor; sus partes principales son tres grandes núcleos subcorticales: caudado, putamen y globo pálido; algunas autoridades también consideran que la amígdala es parte de los ganglios basales (Carpenter, 1991). Estas estructuras están interconectadas y se relacionan funcionalmente con el núcleo subtalámico y la sustancia negra. Junto con el cerebelo, el sistema corticoespinal y los núcleos motores del tallo cerebral, los ganglios basales participan en el control de movimientos. A diferencia de otros componentes del sistema motor, estos últimos no tienen conexiones directas con la médula espinal, y sus funciones motoras son indirectas y se median a través de conexiones neuronales con la corteza cerebral frontal.

El síndrome más común producido por el daño a los ganglios basales es el **mal de Parkinson** (Cote y Crutcher, 1991), en el que se observan tres tipos característicos de trastornos motores: movimiento involuntario, incluyendo temblor; pobreza y lentitud de movimiento, sin parálisis; y cambios en postura y tono muscular. En sus últimas etapas, esta enfermedad se denota por una expresión facial inmóvil, similar a una máscara, dificultad extrema para iniciar movimientos y temblor fino que puede desaparecer una vez que el movimiento se ha iniciado.

Los pacientes con mal de Parkinson también presentan deficiencias cognitivas específicas, lo cual sugiere que los ganglios basales no sólo contribuyen al movimiento, sino también al pensamiento. Las deficiencias observadas incluyen disminución en la flexibilidad cognitiva y una leve reducción en aprendizaje y rememoración (Koss, 1994); también se presenta pérdida de la espontaneidad y falta de iniciativa (La Rue, 1992).

CUERPO CALLOSO

El **cuerpo calloso** es la principal comisura que sirve para integrar las funciones de los dos hemisferios cerebrales. Este gran haz de fibras nerviosas subcorticales tiene aproximadamente 10 cm de longitud y menos de 1 cm de grosor. El cuerpo calloso abarca el cerebro de un lado a otro, justo por encima del nivel del tálamo; aunque existen excepciones, suele conectar sitios cerebrales homólogos en los hemisferios izquierdo y derecho.

Su función se había comprendido mal hasta el decenio de 1960, cuando Sperry, Gazzaniga y otros investigadores iniciaron complejos estudios de laboratorio con los llamados pacientes con escisión hemisférica (Sperry, 1964; Gazzaniga, 1970; Gazzaniga y LeDoux, 1978). Estos pacientes eran personas con epilepsia a quienes se les había escindido el cuerpo calloso para evitar el transporte de las descargas epilépticas de un hemisferio al otro. Aunque exteriormente parecían normales, estos individuos revelaban un sorprendente aislamiento de la conciencia cuando la información visual se restringía a un hemisferio o al otro. Por ejemplo, cuando se presentaba a través de un taquitoscopio, la imagen de una manzana al lado izquierdo del punto de fijación de la persona examinada, el estímulo se procesaba sólo en el hemisferio derecho (debido al cruzamiento normal de las fibras nerviosas). Además, debido a que se había cortado el cuerpo calloso, la imagen de la manzana permanecía atrapada en el hemisferio derecho. Como el lector descubrirá después, en general el hemisferio derecho es mudo y no cumple con funciones importantes del lenguaje. De modo que cuando se les preguntaba a los examinandos "¿Qué vio usted?" quienes respondían desde el hemisferio izquierdo verbal decían honestamente "nada". No obstante, estos pacientes podían identificar con facilidad el objeto señalándolo con la mano izquierda (que está bajo el control neuronal del hemisferio derecho). Esto sugiere que aunque el hemisferio derecho no puede hablar, tiene una capacidad separada e independiente para percibir, aprender, recordar y expedir órdenes para tareas motoras.

En un individuo sano con el cuerpo calloso intacto, la conciencia parece ser unitaria porque las dos mitades del cerebro se pueden comunicar y formar un convenio en cuanto a percepción, pensamiento y acción. Gran parte del conocimiento que se tiene acerca de las especializaciones hemisféricas, que se analizarán después, se ha logrado a partir del estudio detallado de los pacientes con escisión hemisférica. Se ha obtenido mayor discernimiento a partir de estudios sobre personas afectadas por la ausencia congénita de esta estructura, padecimiento que se conoce como agénesis del cuerpo calloso. Por lo general, estos pacientes tienen una variedad de defectos intelectuales, lo que indica que el cuerpo calloso facilita el aprendizaje en muchas esferas cognitivas diferentes (Bradshaw y Mattingley, 1995).

CORTEZA CEREBRAL

La **corteza cerebral**, la capa más exterior del cerebro, es la fuente de los niveles más elevados de procesamiento sensorial, motor y cognitivo. También llamada neocorteza, la corteza cerebral es un desarrollo evolutivo muy reciente. Es la capacidad funcional de este sistema cerebral —con una profundidad uniforme de seis capas— la que divide de manera más trascendental a los humanos de los animales inferiores.

El tejido de la corteza cerebral se encuentra plegado en elaboradas circunvoluciones que consisten en protuberancias y canales. Las protuberancias prominentes se denominan circunvoluciones, mientras las hendiduras, fisuras y canales, cisuras. Esta disposición permite que el cerebro tenga una cantidad mucho mayor de corteza cerebral que si la superficie fuera lisa. Aunque el patrón de circunvoluciones y cisuras es sutilmente único para cada persona, en un cerebro normal siempre pueden discernirse ciertos hitos principales, como la cisura central y la cisura lateral (figura 9–2).

Una pequeña porción de la corteza cerebral es la corteza comprometida. Estos sitios se dedican al procesamiento sensorial básico de la visión, audición, tacto y control motor. No obstante, la especificidad de la corteza comprometida es relativa, no absoluta. Por ejemplo, la circunvolución precentral se considera como la franja motora, sin embargo sólo 40% de las células motoras primarias que ejecutan los movimientos voluntarios se localizan aquí, otro 10 a 20%, en la franja sensorial, y 40 a 50% en sitios cerebrales adyacentes (Brodal, 1981). Lo que es más, la franja motora también contiene

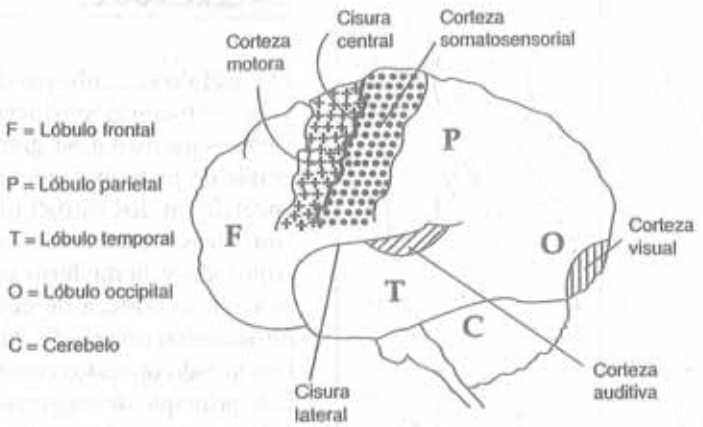


Figura 9–2. Principales hitos del hemisferio cerebral izquierdo.

una proporción de buen tamaño de células sensoriales. Así, las células que satisfacen funciones sensoriales y motoras específicas se encuentran altamente concentradas en el área comprometida respectiva, pero también se van reduciendo y sobreponiendo con sitios cerebrales cercanos.

La mayor parte de la corteza cerebral es corteza no comprometida o de asociación. Estos sitios participan en el análisis de información sensorial y formulación de respuestas motoras. Es la proporción relativamente grande de corteza de asociación lo que distingue al *homo sapiens* de los animales inferiores. El pensamiento abstracto, la creatividad y la solución de problemas se encuentran, en gran medida, inscritas bajo estas adaptaciones evolutivas del cerebro humano.

Benson (1994) subdivide la corteza de asociación en tres tipos: unimodal, heteromodal y supramodal. Las áreas unimodales de asociación se dedican a una sola modalidad (visión, audición, somestesia o movimiento) y se localizan adyacentes a las áreas corticales comprometidas. Dentro de cada área unimodal de asociación, los estímulos que ingresan se categorizan y comparan con la información previa. El daño a las áreas unimodales tiende a causar dificultades en la discriminación y categorización de los estímulos que ingresan dentro de esa modalidad. Por ejemplo, una lesión en el área unimodal visual dejará intacta la visión, pero alterará el reconocimiento y discriminación de estímulos visuales. Se han diseñado diferentes pruebas que implican figuras superpuestas para evaluar el deterioro de este tipo. Por ejemplo, una persona con daño al área unimodal de asociación visual será incapaz de reconocer las formas individuales en la figura 9–3. Las áreas heteromodales de asociación poseen conexiones neuronales que viajan entre modalidades. En estas áreas se entremezcla y procesa la información transmodal en la que participan modalidades sensoriales (visión, audición, tacto). El daño a la corteza heteromodal de asociación puede causar una variedad de síndromes muy conocidos como la afasia, la cual se analizará después. Por último, la corteza supramodal de asociación, localizada exclusivamente en el área frontal, es responsable del control de la cognición de alto nivel a través de un proceso de selección y, en especial, de inhibición. Esta corteza es la más reciente en sentido filogenético, y ejerce control ejecutivo sobre otras funciones cerebrales. El daño a esta área del cerebro puede causar deterioro en el control de

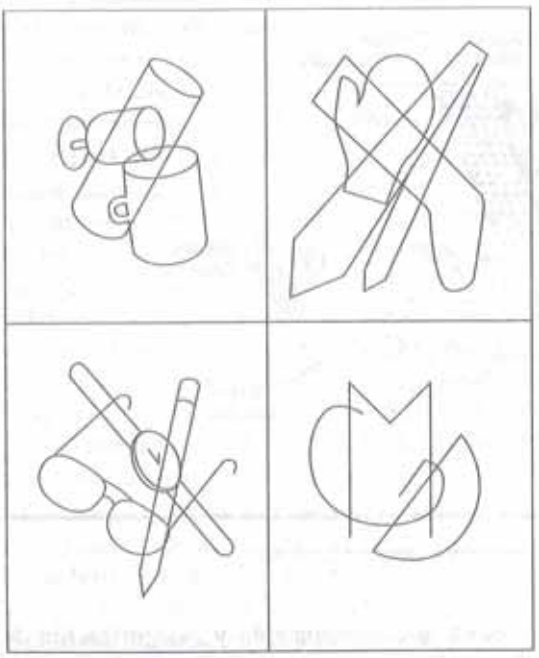


Figura 9–3. Ejemplos de figuras superpuestas para la evaluación de la disfunción visoperceptual.
Fuente: Reproducido con autorización de Gregory, R. J. (1987). *Adult intellectual assessment.* Boston, MA: Allyn & Bacon.

impulsos, en la planificación y de otras funciones ejecutivas.

En lo que resta de esta sección, se examinarán dos aspectos de la función cerebral que ocurren dentro de la corteza cerebral: 1) localización de la función dentro de los cuatro lóbulos principales de cada hemisferio cerebral y 2) lateralización de la función dentro de los hemisferios cerebrales izquierdo y derecho, con atención especial a las funciones de lenguaje del hemisferio izquierdo. También se introducirá al lector a las técnicas de imagenología del cerebro y otras pruebas clínicas de la disfunción cerebral, debido a que, con frecuencia, su utilización va de la mano con la valoración neuropsicológica. Este análisis se diseñó para ayudar al lector a comprender la naturaleza y propósito de las pruebas neuropsicológicas que se analizarán en el siguiente tema.

FUNCIONES DE LOS LÓBULOS CEREBRALES

El encéfalo se conforma de la corteza cerebral y las estructuras subyacentes del telencéfalo. Una cisura media de gran tamaño que se extiende de la parte frontal a la trasera divide al encéfalo en dos estructuras aproximadamente simétricas conocidas como hemisferio cerebral izquierdo y hemisferio cerebral derecho. En general, la corteza de cada hemisferio recibe información sensorial y envía información motora al lado opuesto o contralateral del cuerpo. Este principio de regulación contralateral tiene sólo una excepción real, el sentido del olfato. La estimulación olfatoria en cada fosa nasal se procesa directamente en el bulbo olfatorio del mismo lado que se encuentra justo por encima de la cavidad nasal. También se debe mencionar que la regulación contralateral de la visión es compleja: cada mitad del campo visual se procesa en el hemisferio cerebral opuesto. Lo que es más, la regulación contralateral es relativa, no absoluta. Por ejemplo, las conexiones neuronales del oído derecho se extienden principalmente al hemisferio izquierdo, pero también existen algunas proyecciones hacia el mismo lado o ipsolaterales.

Dentro de cada hemisferio cerebral pueden utilizarse varios hitos prominentes para demarcar cuatro lóbulos principales. El lóbulo occipital se localiza en la parte trasera del cerebro, detrás de la cisura parietoccipital; el lóbulo parietal se encuentra detrás de la cisura central; el lóbulo temporal está debajo de la cisura lateral y el lóbulo frontal se encuentra al frente de la cisura central. Los límites occipitoparietales y occipitotemporales son un tanto indistintos (figura 9–2).

Los lóbulos del mismo nombre en ambos lados del cerebro son aproximadamente simétricos en estructura y también comparten muchas funciones. Por el momento, se enfatizarán las semejanzas operativas de las dos mitades del cerebro a medida que se revisen las funciones de los cuatro lóbulos. No obstante, debe advertirse al lector que los hemisferios cerebrales

izquierdo y derecho también poseen funciones especializadas que se analizarán después. Además, es frecuente que estas especializaciones hemisféricas se correlacionen de manera sensible con variaciones estructurales. Por ejemplo, ciertas estructuras hemisféricas son más grandes en el lado izquierdo del cerebro, lo cual revela la lateralización de las funciones del lenguaje a este hemisferio.

LÓBULOS OCCIPITALES

Las principales áreas sensoriales para la visión se localizan en los **lóbulos occipitales**; gran parte de esta área de proyección se encuentra en la superficie mesial o media que separa los dos hemisferios cerebrales. Cada lóbulo occipital ve el lado opuesto del mundo visual. Así, todos los estímulos visuales del lado izquierdo del punto de fijación del lector se procesan en última instancia en el lóbulo occipital derecho y viceversa. El mundo visual dividido se comparte a través del esplenio, la porción trasera del cuerpo calloso, lo cual produce una percepción unificada del campo visual completo. El daño al área visual principal produce una pérdida correspondiente del campo visual del lado opuesto. Por ejemplo, una lesión amplia en el lóbulo occipital izquierdo haría que una persona fuera ciega del lado derecho. Una lesión muy pequeña podría producir un escotoma o punto ciego.

Un examen minucioso del campo visual es importante para una valoración neuropsicológica completa. Con base en el patrón de pérdida del campo visual, un examinador competente puede inferir la localización y grado del daño cerebral. El sistema visual puede compararse con cables de transmisión que atraviesan el cerebro desde los ojos al frente hasta los lóbulos occipitales atrás. En particular hacia el área trasera del cerebro, los cables irradian al exterior para ocupar porciones importantes de tejido subcortical. Las lesiones subcorticales en la parte trasera del cerebro tienen buena posibilidad de alterar o dañar las redes de transmisión neuronal. Este daño conduce a una pérdida del campo visual asociado (figura 9–4). El patrón de pérdida del campo visual es clave directa del sitio del daño dentro de los lóbulos occipitales y tejido circunvecino.

La porción frontal de cada lóbulo occipital es corteza unimodal de asociación. Estas regiones sintetizan los estímulos visuales y obtienen significados de ellos. Aquí es donde ocurre el procesamiento de alto nivel de la información visual. El daño a la corteza de asociación de los lóbulos occipitales puede provocar **agnosia visual**, una dificultad en el reconocimiento de dibujos, objetos o rostros (Kandel, 1991). Luria (1973) describió el caso típico de un paciente con una lesión de ese tipo:

> El paciente examina cuidadosamente el dibujo de un par de anteojos que se le muestra. Está confundido y no sabe lo que representa. Comienza a adivinar. "Hay un círculo... y otro círculo... y una barra... una barra cruzada... ya veo ¿debe ser una bicicleta?

Las agnosias visuales se vinculan de manera particular con lesiones en el lado derecho de la corteza occipital de asociación, pero también pueden implicar daño en los lóbulos parietal y temporal. Una forma particularmente destacada de agnosia visual es la prosopagnosia, la incapacidad para reconocer rostros familiares. Benson (1994) cita un ejemplo de un hombre de 70 años de edad que había sufrido una serie de embolias que afectaban las porciones delanteras de los lóbulos occipitales. La dolencia principal del paciente era que no podía reconocer a su esposa y a su hija al verlas, aunque las identificaba de inmediato por sus voces. En otro caso de agnosia visual, conocida como agnosia de objeto, un paciente reprodujo con gran habilidad el dibujo de un tren, pero no tenía idea de qué había dibujado. Benson (1988) describe los muchos síntomas fascinantes de la agnosia visual.

LÓBULOS PARIETALES

Kolb y Whishaw (1990) señalan que los lóbulos parietales son un artificio de la definición

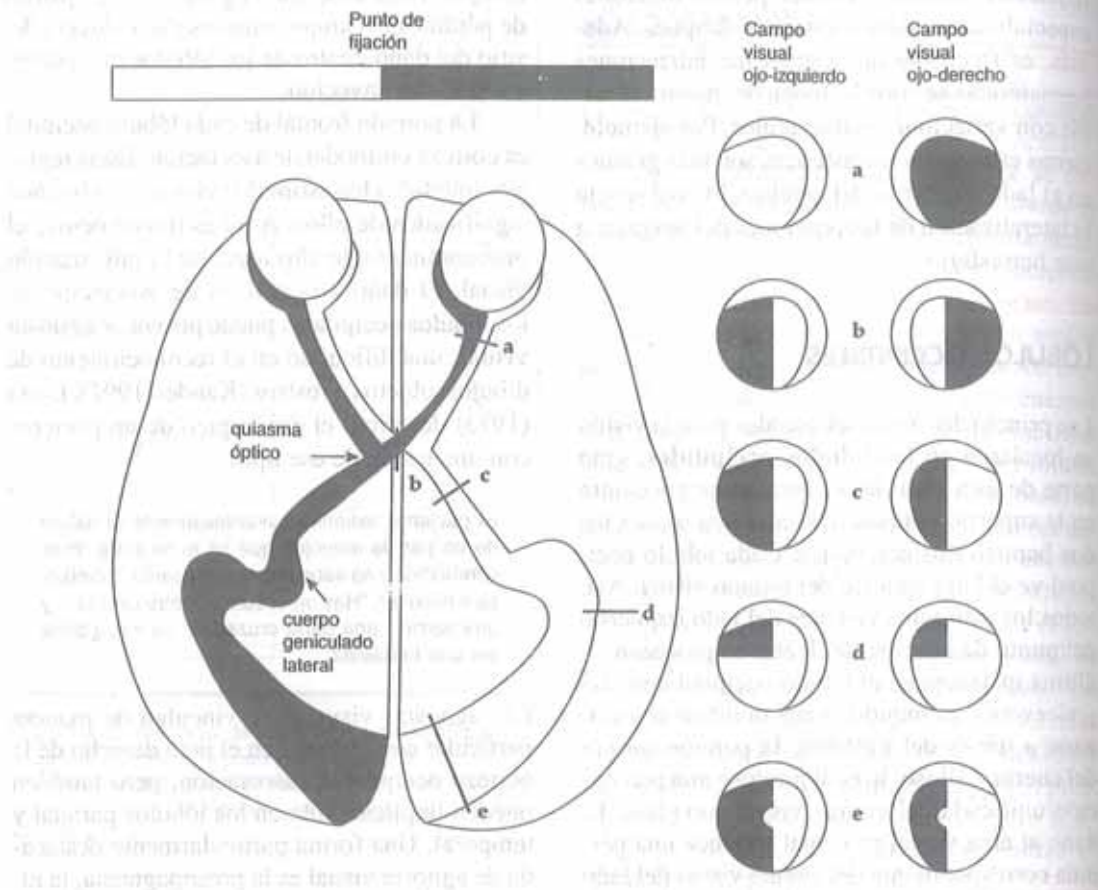


Figura 9–4. Diagrama esquemático del sistema visual y de los efectos de las lesiones en diversos sitios.
Fuente: Adaptado con autorización de Gregory, R. J. (1987). *Adult intellectual assessment.* Boston: Allyn & Bacon.

anatómica exagerada; es decir, no reflejan unidad citoarquitectónica o fisiológica. Por tanto, es imposible una teoría unitaria de la función del lóbulo parietal. De hecho, pueden identificarse dos funciones independientes de los **lóbulos parietales**, una que se ocupa principalmente de la franja somatosensorial localizada en la circunvolución poscentral y la otra que se encarga primordialmente de las zonas traseras o de asociación de los lóbulos parietales.

La circunvolución poscentral de cada lóbulo parietal media la consciencia de lo que sucede en la superficie contralateral del cuerpo. Este tejido, al que de modo apropiado se le denomina franja somatosensorial (o somestésica), recibe información sobre las sensaciones del tacto y posición de los miembros (figura 9–2). La franja somatosensorial y la franja motora del lóbulo frontal se encuentran muy interrelacionadas por redes neuronales subcorticales; la organización tópica de estos sistemas funcionales entrelazados es muy similar. La cantidad de tejido somatosensorial que se dedica a cada lugar del cuerpo corresponde a la importancia funcional de esa área. Por ejemplo, la realimentación de los labios y lengua es esencial para el habla; la cantidad de tejido somatosensorial que se dedica a estas estructuras es relativamente grande. El daño a la franja somatosensorial resulta en pérdida de la sensación

en las partes correspondientes del cuerpo en el lado opuesto.

La porción trasera de cada lóbulo parietal es corteza heteromodal de asociación. Esta área parece haberse especializado en la integración de información sensorial de regiones somáticas, visuales y auditivas cercanas. Un síntoma común del daño a esta área es la incapacidad para reconocer objetos a través del tacto, síntoma que se conoce como astereognosis. Muchas baterías de pruebas neuropsicológicas incorporan medidas de astereognosis, como el reconocimiento de monedas (1 centavo, 5 centavos, 10 centavos) a través del tacto. Ésta es una tarea transmodal, en la medida en que el paciente debe formar una representación visual basándose sólo en el tacto. Esta región utiliza su información polimodal para proporcionar guía visomotora a los miembros, manos y ojos (Clarke, 1994).

El daño a la región parietal de asociación en cualesquiera de los dos lados puede alterar la capacidad para el dibujo. Ésta es una forma de dispraxia de construcción que se analizará posteriormente. El daño a esta área en el lado izquierdo reduce la calidad del dibujo, como si la persona tuviera dificultad para lograr que la mano que realiza el dibujo vaya en dirección correcta. En contraste, el daño a esta área en el lado derecho resulta en déficit perceptual. La persona tiene problemas para integrar las partes individuales en un todo consistente; se pierde la *gestalt* general del dibujo. Las áreas de asociación de los lóbulos parietales izquierdo y derecho también cumplen con varias funciones lateralizadas, las cuales se analizarán después.

LÓBULOS TEMPORALES

Se pueden identificar cuatro funciones básicas de los **lóbulos temporales** (Kolb y Whishaw, 1990; Bear, 1986): procesamiento primario de sensaciones auditivas, procesamiento secundario de percepciones auditivas, almacenamiento de memoria a largo plazo y modulación de impulsos biológicos (en especial agresión, temor y sexualidad). Cada una de estas funciones se analizarán por separado.

Cada lóbulo temporal contiene una corteza auditiva primaria en su porción superior. Gran parte de la corteza auditiva está insertada en la cisura lateral y no es visible desde una perspectiva lateral del cerebro. La corteza auditiva izquierda recibe información sensorial que proviene principalmente del oído derecho y viceversa; sin embargo, existen también algunas entradas sensoriales del mismo lado. Una lesión a la región auditiva primaria en un lóbulo temporal puede causar una variedad de deficiencias audioperceptuales, incluyendo incapacidad para detectar sonidos breves, deterioro en la localización de los mismos y dificultad para discriminar los sonidos del habla. Estas deficiencias son especialmente probables en el caso de daño en la corteza auditiva izquierda, como se verá más adelante. En casos aislados, el daño a las áreas auditivas puede causar sordera, aunque el aparato auditivo esté intacto.

La corteza auditiva unimodal de asociación ocupa gran parte de la superficie externa del lóbulo temporal. Esta corteza es importante para el análisis de sonidos complejos y estructuras rítmicas acústicas. Las regiones auditivas de asociación de los lóbulos temporales izquierdo y derecho se especializan en diferentes funciones que se analizarán posteriormente. En pocas palabras, el lado izquierdo gobierna la comprensión del lenguaje, mientras el derecho los patrones y ritmos auditivos no verbales. El daño a esta región en el lóbulo temporal izquierdo altera la comprensión del lenguaje, mientras el daño al lado derecho origina agnosia auditiva, dificultad para reconocer y discriminar sonidos no verbales. Albert, Sparks, von Stockert y Sax (1972) informaron el fascinante caso de un hombre de 57 años que sufrió una embolia aislada en la corteza de asociación del lóbulo temporal derecho. Éste, quien en otros sentidos tenía excelente funcionamiento, podía notar la existencia de un sonido, pero no podía distinguir el timbre del teléfono, el ladrido de un perro, el golpeteo de los cascos de un caballo, un silbido, el palmoteo con las manos o el chasquido de los dedos. No obstante, su percepción de los sonidos del lenguaje (una función del lóbulo temporal izquierdo) estaba intacta.

Junto con el hipocampo, las porciones más centrales de la corteza temporal ayudan a mantener las funciones de memoria a largo plazo. Como se estudia más adelante, estas funciones son muy asimétricas, con las funciones verbales a la izquierda y las pictóricas a la derecha. En pocas palabras, las lesiones al lóbulo temporal izquierdo, a la estructura subyacente del hipocampo o a ambas, deterioran el recuerdo demorado de material verbal, como párrafos o listas de palabras, ya sea que se presente de manera visual u oral. Las lesiones al lóbulo temporal derecho, a la estructura subyacente del hipocampo o a ambos causan defectos en el recuerdo demorado de material pictórico, como dibujos geométricos y rostros. Es típico que las lesiones al lóbulo temporal no alteren el recuerdo **inmediato** de material verbal o no verbal.

Por último, junto con las estructuras subyacentes de la amígdala, los lóbulos temporales también participan en impulsos biológicos como agresión, temor y sexualidad. La evidencia para la participación de los lóbulos temporales en la motivación y emoción proviene de dos fuentes: estudios sobre estimulación eléctrica directa de esta región e investigación sobre la alteración conductual en personas con trastornos convulsivos del lóbulo temporal. Por ejemplo, Penfield y Jasper (1959) informaron que la estimulación eléctrica leve de la corteza frontal y media del lóbulo temporal producían sentimientos de temor, una respuesta que también se obtiene de la amígdala. Bear (1986) ha catalogado los trastornos conductuales que pueden ser resultado de la disfunción emocional secundaria al daño al lóbulo temporal (cuadro 9–3). Estos síntomas describen la llamada personalidad del lóbulo temporal, aunque pocas personas combinan todos estos rasgos.

Se cierra este análisis con una nota precautoria. Debido a que estos lóbulos son ricos en conexiones subcorticales hacia los lóbulos parietal, frontal, límbico y occipital del cerebro, las lesiones en esta región pueden causar síntomas diversos que no se catalogan en lo dicho antes. Por ejemplo, las lesiones al temporal pueden interferir con el reconocimiento visual. Esto lo demostró inicialmente Milner (1958), quien encontró que sus pacientes tenían

Cuadro 9–3. Resumen de síntomas que pueden ocurrir en la disfunción del lóbulo temporal

Compulsividad e hipersexualidad indiscriminada
Hiperirritabilidad ante insignificancias triviales
Ansiedad y respuestas fóbicas
Preocupaciones paranoides que se generalizan de manera amplia
Periodos de depresión, de euforia con agitación o ambos
Preocupación por la religión, cosmología o filosofía
Escritura, dibujo o disertaciones extensas, pero improductivas
Preocupación por los detalles
Circunstancialidad, un estilo locuaz y tortuoso
Viscosidad, una tendencia a prolongar los encuentros sociales

Fuente: Basado en Bear, D. M. (1986) Behavioural changes in temporal lobe epilepsy. En M. R. Trimble y T. G. Bolwig (Eds.) *Aspects of epilepsy and psychiatry.* Nueva York: John Wiley.

un deterioro en el reconocimiento de anomalías visuales (p, ej., un mono dentro de una jaula con una pintura al óleo sobre la pared). Kolb, Milner y Taylor (1983) encontraron que los pacientes con lesiones al temporal derecho no podían percibir la porción de un rostro del lado del campo visual izquierdo. Kolb y Whishaw (1985) describen este interesante síntoma:

> ... los pacientes con daño al lóbulo temporal derecho no parecen capaces de percibir las señales sociales sutiles como esas miradas discretas, pero obvias, al propio reloj, un ademán que con frecuencia tiene el propósito de indicar que se debe suspender una conversación. Se puede suponer que los pacientes no son capaces de percibir el significado de la señal visual.

El lector interesado puede consultar a Kolb y Whishaw (1990) para un examen más amplio de la disfunción del lóbulo temporal.

LÓBULOS FRONTALES

Los **lóbulos frontales** se requieren para la programación, regulación y verificación de las fun-

ciones ejecutivas y el desempeño motor (Luria, 1973; Lezak, 1995). Las funciones ejecutivas incluyen formulación y planificación de metas; ejecución de planes dirigidos al logro de objetivos y desempeño eficiente. Es con los lóbulos frontales que los humanos intentan, forman planes y regulan su conducta al comparar los efectos de sus actos con las intenciones originales.

La puesta en marcha de un plan requiere de un movimiento corporal de cierto tipo. Las personas buscan alcanzar sus metas a través de la manipulación física del ambiente, ya sea con sus manos o a través de la actividad motora del habla. Por tanto, no es sorprendente encontrar que la corteza motora principal se localiza en los lóbulos frontales —donde también se forman los planes e intenciones.

La corteza motora primaria se encuentra en la circunvolución precentral, en la parte trasera del lóbulo frontal, justo enfrente de la cisura central. La corteza motora tiene funciones contralaterales, donde la corteza motora izquierda controla los movimientos corporales a la derecha y viceversa. Penfield (1958) fue el primero en establecer un mapa de la organización tópica de la franja motora durante una serie de operaciones para retirar el tejido cortical dañado en personas con epilepsia. Este investigador estimuló diferentes áreas de la corteza motora con una corriente eléctrica inofensiva para obtener un mapa de la correspondencia entre aquélla y diferentes partes del cuerpo. Penfield encontró que aquellas áreas del cuerpo que requieren un control preciso, como dedos y boca, ocupan una cantidad desmesuradamente grande de espacio cortical.

Justo al frente de la corteza motora primaria se encuentra la corteza motora complementaria. Ésta participa en el ordenamiento en serie de complejas cadenas motoras; es decir, en la programación del movimiento. Una porción de los lóbulos frontales justo por debajo de la corteza motora complementaria participa en el control de la mirada voluntaria. El lóbulo frontal izquierdo también media el lenguaje expresivo, que se examina en detalle más adelante.

El daño a la corteza motora primaria causa problemas en el lado contrario del cuerpo en cuanto a control motor fino y también reduce la velocidad y fortaleza de los movimientos de los miembros. Estos efectos se detectan con facilidad con pruebas motoras sencillas como velocidad de golpeteo con los dedos. El daño grave a la corteza motora causa parálisis total de las partes corporales afectadas. El daño a la corteza motora complementaria causa deficiencias en la ejecución de secuencias motoras como la imitación de una serie de movimientos faciales y con el brazo (Kolb y Milner, 1981).

La causa más común de daño al lóbulo frontal es la lesión cefálica cerrada, que es un tipo de lesión traumática de la cabeza (en el cuadro 9–4 se describe este tipo de lesión y otras causas destacadas de neuropatología en el adulto, incluyendo síndromes relacionados con la edad). En una lesión cefálica cerrada, se aplican de manera instantánea fuerzas de aceleración/desaceleración a todo el cerebro, como cuando la cabeza de una persona choca contra el tablero de un automóvil en un accidente. Debido a las superficies irregulares del cráneo circundante, casi siempre resultan dañadas las superficies inferoanteriores de los lóbulos frontales (Jennett y Teasdale, 1981). Los extremos frontales de los lóbulos temporales también son sumamente vulnerables en la lesión cefálica cerrada.

Nauta (1971) resume los efectos de la disfunción del lóbulo frontal como un "trastorno de la programación conductual". Lezak (1983, 1995) ha catalogado los trastornos conductuales que pueden resultar del daño generalizado, bilateral, al lóbulo frontal:

1. Problemas de tipo motivacional que implican disminución de la espontaneidad y productividad, frecuencia reducida de la conducta y falta de iniciativa.
2. Dificultades para hacer cambios mentales y perseverar en actividades y respuestas.
3. Problemas para reprimirse, que con frecuencia se describe como impulsividad, reactividad exagerada y dificultad para contener una respuesta incorrecta o indeseable.
4. Deficiencias en la autoconciencia que resultan en incapacidad para percibir los errores en el desempeño o estimar de manera apropiada las situaciones sociales.

Cuadro 9–4. Principales trastornos neuropatológicos de la adultez y el envejecimiento: sinopsis breve y referencias esenciales

Lesión cerebral traumática

Bingler (1990), Dikmen y colaboradores (1995)

Descripción: las consecuencias neurológicas dependen de la gravedad de la lesión, pero todas las siguientes son posibles: contusiones o magulladura del cerebro bajo el sitio del impacto (lesión en golpe); contusiones en el lado opuesto (lesión en contragolpe); contusiones frecuentes en las superficies inferiores de los lóbulos frontales y en los extremos de los lóbulos temporales; lesión axonal difusa o daño inespecífico debido a efectos de fuerza de corte-distensión sobre las vías neuronales; daño al tejido cerebral debido a obstrucción del flujo sanguíneo; hematoma o coágulo de sangre entre el cráneo y la superficie del cerebro; edema o inflamación del mismo; las consecuencias a largo plazo incluyen la posible disminución en el tamaño del cerebro y el agrandamiento correspondiente del sistema ventricular.

Efectos neuroconductuales potenciales: las dolencias más comunes, y confiables, se refieren a problemas de concentración y memoria; otras generalizaciones son difíciles debido a que la naturaleza y gravedad del daño cerebral no será el mismo en cualesquiera de dos pacientes; el daño focal puede conducir a síntomas específicos; por ejemplo, el daño en las áreas de lenguaje del hemisferio izquierdo puede causar afasia expresiva; muchos estudios sugieren que los pacientes con lesión cerebral traumática tienen una afectación más grave a causa de los trastornos de personalidad y emocionales que de las discapacidades cognitivas y físicas (Lezak y O'Brien, 1990).

Enfermedad neoplásica (Tumor)

Reitan y Wolfson (1993)

Descripción: la enfermedad neoplásica engloba muchas formas diferentes de crecimientos tumorales; por ejemplo, los gliomas son tumores en forma de zarcillo de células gliales que infiltran al cerebro a lo largo de un periodo de semanas o meses; los mengiomas son tumores de crecimiento más lento, con forma globular, de las meninges (las membranas que encierran al cerebro), las cuales presionan el cerebro.

Efectos neuroconductuales potenciales: los tumores cerebrales producen una variedad de efectos, dependiendo de la localización y tamaño; los tumores de infiltración rápida pueden comprometer muchas habilidades; por ejemplo, habilidades de lenguaje y para solución de problemas, funciones motoras y sensoriales del lado derecho, si está afectado el hemisferio izquierdo; los mengiomas de crecimiento más lento pueden conducir a síntomas focales que se relacionan con el sitio de invasión en el cerebro; por ejemplo, déficit en la capacidad espacial y deterioro en las funciones motoras y sensoriales en el lado izquierdo si se afecta al área parietofrontal derecha.

Abuso crónico del alcohol

Dávila y colaboradores (1994)

Descripción: la ingestión crónica de alcohol conduce a cambios en las neuronas que incluyen pérdida de ramas y espinas dendríticas, en especial, en el hipocampo y en la circunvolución dentada; también se observa agrandamiento de los ventrículos y aumento en la amplitud de las cisuras; en casos graves se encuentra atrofia del tálamo medial y de los cuerpos mamilares (síndrome de Wernicke-Korsakoff); la neuropatología del alcoholismo puede exacerbarse por deficiencias vitamínicas y nutricionales.

Efectos neuroconductuales potenciales: en casos de abuso grave del alcohol en el que se encuentran comprometidos el tálamo medial y los cuerpos mamilares, se observa la profunda amnesia anterógrada del síndrome de Wernicke-Korsakoff; los pacientes muestran una incapacidad para conservar el recuerdo de acontecimientos por más de un corto tiempo, aunque la memoria inmediata está intacta y la memoria remota sólo muestra un daño leve; se presenta confabulación o falsificación de recuerdos con conciencia clara; otros síntomas de abuso grave incluyen alteraciones de la marcha y dificultades con la mirada; en alcohólicos neurológicamente intactos, los efectos neuroconductuales son más esquivos, pero pueden incluir deficiencias sutiles en memoria y dificultades para la solución de problemas novedosos.

Enfermedad de Alzheimer

Knight (1992), Koss (1994)

Descripción: la enfermedad neurológica degenerativa más común es la Enfermedad de

Cuadro 9–4. Principales trastornos neuropatológicos de la adultez y el envejecimiento: sinopsis breve y referencias esenciales (continuación)

Alzheimer (EA), que presenta una degeneración insidiosa del cerebro; la fisiopatología incluye depósitos con forma de masas en el cerebro que constan de placas neuríticas y marañas neurofibrilares; cambios cerebrales adicionales incluyen pérdida neuronal, encogimiento o atrofia del cerebro, reducción de los neurotrasmisores acetilcolina que participan en la memoria y acumulación de depósitos extraños en la vasculatura cerebral; el curso invariable de la enfermedad es en declive. Descrito inicialmente en 1907, Alois Alzheimer describió su primer caso como sigue:

> El primer síntoma notable de la enfermedad que mostró esta mujer de 51 años de edad fue suspicacia acerca de su esposo. Pronto se hizo evidente un aumento rápido de su deterioro de memoria; ya no podía orientarse en su propia vivienda, llevaba los objetos de un lado a otro y los ocultaba y en ocasiones, al creer que otras personas pretendían asesinarla, comenzaba a gritar a voz en cuello. Al observarla en la institución, todo su proceder tiene el aspecto de completa perplejidad. Está absolutamente desorientada en tiempo y lugar (La Rue, 1992).

Aunque la enfermedad de Alzheimer no es parte del envejecimiento normal, la edad avanzada es un factor importante de riesgo; rara antes de los 65 años de edad, la enfermedad aflige a 3% de las personas entre 65 y 74 años, a 18% de aquéllas entre 75 y 84 años y a casi la mitad de aquellos de 85 años en adelante (Evans, Funkenstein, Albert y otros, 1989).

Efectos neuropsicológicos potenciales: como detallaron Storandt y Hill (1989), la dificultad para la adquisición de información nueva (disfunción de la memoria a corto plazo) es, en términos generales, el síntoma más notable en las primeras etapas; los pacientes también pueden mostrar una notable disfunción del lenguaje (p. ej., dificultad pronunciada para encontrar las palabras) o un trastorno visoespacial sobresaliente; también son comunes los informes de cambios de personalidad, incluyendo delirios y agitación; las últimas etapas se caracterizan por una discapacidad grave y generalizada.

Demencia vascular
Mirsen y Hachinski (1988)

Descripción: segunda causa más común de demencia en los ancianos, la demencia vascular es producto del bloqueo de una arteria y de la consiguiente necrosis del tejido cerebral debido a insuficiencia en la irrigación sanguínea (infarto) o al sangrado dentro o alrededor del cerebro (hemorragia); el inicio repentino es la regla, pero la acumulación de pequeñas embolias a lo largo del tiempo, lo cual se conoce como demencia por infartos múltiples (DIM), puede producir un trastorno aparentemente progresivo. La *Hachinski Ischemic Score* (Puntuación isquémica de Hachinski) se desarrolló para discriminar entre la demencia por infartos múltiples y la enfermedad de Alzheimer (Hachinski, Iliff, Zilha y otros, 1975). La DIM se indica por la presencia de varios de los siguientes factores: inicio abrupto, malestares somáticos, deterioro por pasos, incontinencia emocional, fluctuación en el curso de la enfermedad, antecedentes de hipertensión, confusión nocturna, antecedentes de embolia, preservación de la personalidad, presencia de ateroesclerosis, depresión y signos neurológicos focales.

Debido a que la DIM puede tratarse de manera parcial, el diagnóstico diferencial de ésta contra la EA es más que una cuestión académica; el curso de la enfermedad en la DIM es más corto que en la EA, pero más variable de una persona a otra.

Efectos neurológicos potenciales: el síndrome por accidentes cerebrovasculares se define por el inicio agudo de un déficit focal que implica al Sistema Nervioso Central; los síntomas dependen del sitio del infarto, pero pueden incluir debilidad motora y deterioro en la sensibilidad en los miembros opuestos; si el hemisferio dominante está afectado puede haber presencia de afasia de falta de fluidez; la embolia en la parte trasera del cerebro puede producir pérdida parcial del campo visual; los síntomas agudos pueden disminuir parcialmente y conducir a una meseta de funcionamiento estable.

Mal de Parkinson
La Rue (1992)

Descripción: el mal de Parkinson es casi inexistente antes de los 40 años de edad y afecta

Cuadro 9–4. Principales trastornos neuropatológicos de la adultez y el envejecimiento: sinopsis breve y referencias esenciales (continuación)

sólo a 1 o 2 personas de cada 1 000 con edades de 70 años en adelante; es principalmente un trastorno del movimiento, pero también son comunes los problemas cognitivos y emocionales. Las etapas finales de este mal pueden implicar una demencia evidente; los síntomas incluyen lentitud de movimientos (bradicinesia) temblor en reposo, andar pesado y rigidez postural; la neuropatología implica disminución de dopamina y pérdida neuronal en los ganglios basales.

Efectos neuroconductuales potenciales: el temblor es el síntoma inicial más común y menos debilitante; la tasa de progreso es bastante variable, pero la discapacidad de movimiento en el mal de Parkinson puede volverse pronunciada y conducir al internamiento; de 10 a 20% de los pacientes con este mal desarrollan una demencia evidente; además revelan un déficit en pruebas neuropsicológicas que requieren velocidad (p. ej., Dígitos y símbolos, tareas de trazado de un camino y medidas de tiempo de reacción); de manera sorprendente, las pruebas de discriminación visual y de aprendizaje de pares asociados que no requieren velocidad también diferencian a los pacientes con mal de Parkinson moderado a grave de los controles equiparados (Pirozzolo, Hansch, Mortimer, Webster y Kuskowski, 1982); cerca de 40 a 60% de los pacientes con mal de Parkinson también experimentan depresión.

Síndrome demencial de la depresión
Blazer (1993)

Descripción: cerca de 10 a 20% de los ancianos con depresión muestran deficiencias cognitivas que se asemejan a la demencia orgánica; la pérdida de memoria en combinación con quejas graves de discapacidad son aspectos comunes. El síndrome demencial de la depresión (conocido también como seudodemencia) es principalmente un diagnóstico posterior que se define por un regreso al desempeño cognitivo normal después de que se trata la depresión grave.

Efectos neuroconductuales potenciales: es común la pérdida de memoria tanto de acontecimientos recientes como remotos; se conservan la atención y concentración; las habilidades sociales pueden mostrar pérdidas notables en etapas iniciales; se observa una notable variabilidad en el desempeño en pruebas cognitivas, incluso entre tareas similares.

5. Una actitud concreta (Goldstein, 1944) en la que los objetos, experiencias y conducta se toman por su valor superficial más obvio.

Es curioso que las lesiones al lóbulo frontal pueden tener poco efecto sobre el aprendizaje anterior o las habilidades bien establecidas. Tanto Hebb como Penfield informaron que la extirpación quirúrgica del tejido del lóbulo frontal causaba poco cambio en las puntuaciones del CI (Hebb, 1939; Penfield y Evans, 1935). Los primeros estudios sobre la lobotomía prefrontal demostraron, en gran medida, las mismas conclusiones: no existe cambio en el CI o incluso una pequeña mejoría después de la desconexión de los lóbulos frontales.

Ha resultado difícil el diseño de medidas adecuadas de la función del lóbulo frontal. Lezak (1995) señala que los trastornos de éste cambian la manera en que una persona responde, mientras la mayor parte de las pruebas mide lo que una persona sabe. Esta autora ha diseñado un ingenioso método llamado Prueba *Tinkertoy*®, el cual se analizará en el siguiente tema, para evaluar las dificultades de programación que experimentan las personas con lesiones en el lóbulo frontal. Es más común que los clínicos dependan de la observación y de listas de verificación para diagnosticar la disfunción del lóbulo frontal. Un instrumento útil para este propósito es la *Checklist of Executive Functions* (Lista de Verificación de Funciones Ejecutivas; Pollens, McBratnie y Burton, 1988). Los reactivos en esta lista definen importantes aspectos de las funciones del lóbulo frontal: conciencia, establecimiento de metas, planificación, autoiniciación, autoinhibición, automonitoreo, capacidad para cambiar

de propensión y conducta estratégica. Cada reactivo se califica en una escala de 1 a 5, de modo que la puntuación total proporciona un índice de la integridad de las funciones ejecutivas.

LATERALIZACIÓN CEREBRAL DE LA FUNCIÓN

Hasta este momento se han destacado las semejanzas entre ambas mitades del cerebro, en particular con respecto al procesamiento contralateral de las funciones sensoriales y motoras. No obstante, en muchos aspectos importantes, los dos hemisferios del cerebro humano son asimétricos anatómica y funcionalmente (Cytowic, 1996; Geschwind y Galaburda, 1987). Las asimetrías estructurales importantes incluyen las siguientes:

1. La parte superior del lóbulo temporal, el plano temporal, es más grande en el hemisferio izquierdo.
2. El hemisferio izquierdo contiene más materia gris, aunque es un poco más pequeña y ligera que en el hemisferio derecho.
3. La cisura lateral es mucho más larga en el hemisferio derecho, por lo que la corteza parietotemporal es ligeramente más grande en este lado.

Por supuesto, ocurrirán excepciones en casos individuales. Las diferencias antes listadas son las más consistentes y pronunciadas en personas diestras y en varones. Los zurdos y las mujeres muestran menores asimetrías estructurales y funcionales del cerebro (Weekes, 1994).

Las asimetrías estructurales entre los hemisferios se correlacionan con diferencias funcionales bien conocidas entre ambos lados del cerebro. El mayor tamaño de ciertas secciones del área temporal en el lado izquierdo está, en parte, al servicio del lenguaje, mientras que el de algunas secciones del área parietotemporal en el lado derecho está parcialmente al servicio del pensamiento espacial. En lo que resta de esta sección se catalogarán aún más las especialidades de los hemisferios izquierdo y derecho, pero se advierte de antemano al lector que la lateralización de la función es una cuestión relativa y no absoluta. Por ejemplo, ambos hemisferios tienen cierto grado de capacidad verbal y espacial. Lo que es más, casi cualquier actividad intelectual de alto nivel requiere de la interacción sintética de todo el cerebro (Efron, 1990). El habla es un ejemplo a propósito; aunque es principalmente una función del hemisferio izquierdo, el hemisferio cerebral derecho proporciona los patrones de entonación. Como resultado, los pacientes con lesiones en el lado derecho (en particular en el área frontal) pueden hablar de manera inquietantemente monótona (Gardner, 1975).

EL LENGUAJE COMO FUNCIÓN DEL HEMISFERIO IZQUIERDO

El lenguaje es principalmente (pero no exclusiva) una función del hemisferio izquierdo en la que participan estructuras corticales y subcorticales ampliamente separadas. Debido a que tantas regiones de este hemisferio contribuyen al lenguaje, casi cualquier lesión importante en esta área producirá algún tipo de alteración en su producción o comprensión. Por esta razón, un perfil detallado del lenguaje ofrece una ventana hacia la integridad y funcionamiento del hemisferio izquierdo.

Los conceptos modernos acerca de las correlaciones cerebro-lenguaje provienen, de hecho, del siglo XIX. En 1861, Paul Broca observó que el daño en una pequeña región justo enfrente de la corteza motora del hemisferio izquierdo causaba un trastorno del lenguaje, al que originalmente se llamó afasia expresiva, ahora conocido como afasia de falta de fluidez. Las personas con daño en esta área premotora en el hemisferio izquierdo —llamada como debería ser área de Broca— hablan de manera lenta y trabajosa. Tienen dificultad para enunciar correctamente las palabras; el acto de hablar les parece tortuoso. El habla asume una naturaleza francamente telegráfica; con frecuencia se omiten los adjetivos, adverbios, ar-

tículos y conjunciones —las palabras que le dan color al habla. La redacción también es difícil para estos individuos. Por fortuna, las personas que experimentan afasia de Broca tienen poca dificultad para comprender ya sea el lenguaje hablado o escrito. En su forma pura, este trastorno implica sólo al lenguaje expresivo.

En 1874, Wernicke anunció que el daño a la porción superior y posterior del lóbulo temporal izquierdo —una región que ahora se conoce como área de Wernicke— se vinculaba con un trastorno del lenguaje que en principio se denominó afasia receptiva y ahora se conoce más como afasia fluida. Los individuos afectados parecen incapaces de comprender el lenguaje hablado o escrito. En apariencia, las personas con afasia de Wernicke no tienen dificultad para percibir las palabras, pero no pueden asociarlas con su significado subyacente. En consecuencia, las expresiones escritas y verbales de las personas con esta afasia son fluidas, pero carecen de significado. Por ejemplo, cuando se le pide que defina la palabra **libro**, un paciente podría responder: "libro, un macinto, un rey de preparador, lo encuentras al frente de un auto listo para dirigirlo". La misma persona podría definir **espantapájaros** como: "diremos que es una bruja de riesco de tres minutos, encontrarás una en el campo en tres brujas" (Williams, 1979).

Con base en las observaciones de Broca y Wernicke, Geschwind (1972) propuso un modelo estructural, neurológico de las funciones de lenguaje del hemisferio izquierdo, que ha tenido gran influencia en la evaluación neuropsicológica. Este modelo tiene que ver de manera directa con la evaluación de las habilidades de lenguaje; a continuación se describen los principales elementos y se presentan en la figura 9–5. Geshwind propuso lo siguiente:

1. El lenguaje hablado se percibe en la corteza auditiva izquierda, en la parte superior del lóbulo temporal, y después se transfiere al área de Wernicke.
2. En el área de Wernicke se activan los significados de las palabras y se transportan los códigos auditivos a un haz subcortical de fibras transmisoras llamadas fascículo arqueado.
3. El fascículo arqueado envía los códigos auditivos directamente al área de Broca.
4. Al llegar al área de Broca, el código auditivo activa al código de articulación correspondiente, el cual especifica la secuencia de acciones musculares requeridas para pronunciar una palabra.
5. A su vez, el código de articulación se transmite a las porciones de la corteza motora que gobiernan a la lengua, labios, laringe y así

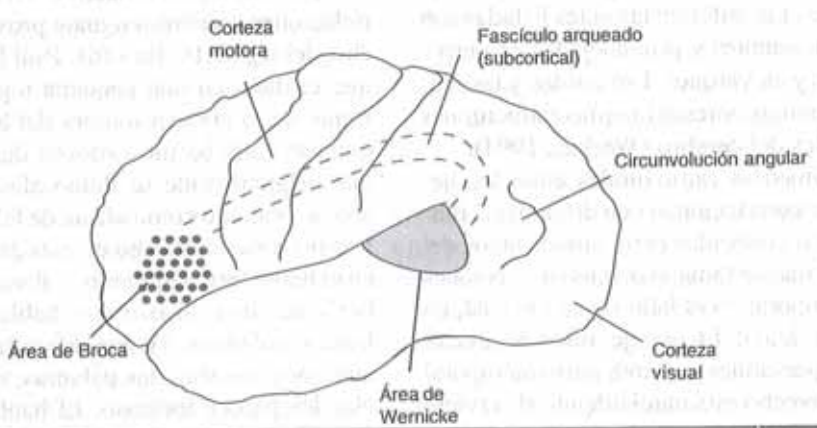


Figura 9–5. Modelo estructural de las funciones de lenguaje del hemisferio izquierdo.

sucesivamente, a fin de producir la palabra hablada que se desea.

Comprender o leer una palabra escrita implica la mayor parte de las vías descritas, pero con un punto de inicio diferente:

6. Las palabras escritas se registran primero en la corteza visual y después se envían a través de la corteza visual de asociación a la circunvolución angular.
7. En la circunvolución angular, la forma visual de la palabra se convierte en un código auditivo que se almacena en el área de Wernicke, con lo cual se tiene acceso al significado de la palabra escrita, que también puede ser hablada (pasos 2 a 5 anteriores).

El modelo de Geshwind es útil para explicar varios síndromes clínicos producidos por el daño cerebral aislado en el hemisferio izquierdo (Gregory, 1999):

- Las lesiones en el área de Broca causarán un habla lenta, trabajosa, telegráfica, pero no se verá afectada la comprensión del lenguaje hablado o escrito.
- El daño al área de Wernicke tendrá implicaciones más graves y generalizadas para la comprensión del lenguaje; es decir, el paciente no podrá comprender las comunicaciones habladas o escritas.
- El daño a la circunvolución angular producirá grave discapacidad para la lectura, pero habrá pocos problemas para comprender el habla o para hablar.
- El deterioro limitado a la corteza auditiva izquierda resultará en grave trastorno de la comprensión verbal. Sin embargo, las personas que lo padezcan podrán hablar y leer con normalidad.

En la práctica, pocos pacientes revelan síntomas de afasia que se encuentren de manera clara dentro de una u otra de las categorías anteriores. Lo que es más, los conceptos modernos sobre la afasia señalan debilidades en el modelo típico (p. ej., su punto de vista demasiado simplista acerca de la estructura del lenguaje) y proponen un modelo complejo, no lineal, de la afasia que se encuentra más allá de la cobertura de este libro (Cytowic, 1996; Whitaker y Kahn, 1994). No obstante, una evaluación minuciosa de las funciones del lenguaje es parte esencial de toda valoración neuropsicológica y el modelo característico de Broca, Wernicke y Geshwind proporciona un punto de inicio provechoso. Pueden encontrarse perspectivas adicionales sobre la afasia y el modelo estructural del lenguaje en Benson (1994) y Mayeux y Kandel (1991).

FUNCIONES ESPECIALIZADAS DEL HEMISFERIO DERECHO

Con base en estudios sobre personas sanas y con daño cerebral, ahora se encuentra ampliamente establecido que el hemisferio derecho es dominante para una variedad de habilidades cognitivas y perceptuales. Sin embargo, un análisis detallado de las funciones especializadas del hemisferio derecho está más allá del espectro de esta sección. Pueden encontrarse reseñas competentes de la extensa literatura sobre este tema en Bradshaw y Mattingley (1995), Joseph (1988), Kupfermann (1991b), Dean (1986) y Springer y Deutsch (1985). En general, el hemisferio derecho parece ser dominante en el análisis del espacio geométrico y visual, comprensión y expresión de emociones, procesamiento de la música y sonidos ambientales no verbales, producción de recuerdos no verbales y espaciales y reconocimiento táctil de formas complejas.

Un síntoma frecuente de daño en el hemisferio derecho es la **dispraxia de construcción**, que es el deterioro en la capacidad para manejar relaciones espaciales en un esquema bidimensional o tridimensional (Reitan y Wolfson, 1993). Este síntoma suele exhibirse a través del deterioro en la capacidad para copiar formas sencillas como una cruz. Las lesiones al hemisferio izquierdo también pueden causar dispraxia de construcción, pero la correlación es menos

consistente. La mayor parte de las baterías de pruebas neuropsicológicas incluyen una o más tareas de copiado para detectar este trastorno. En el cuadro 9–5 se resumen las conclusiones sobre lateralización cerebral.

PRUEBAS CLÍNICAS Y TÉCNICAS DE IMAGENOLOGÍA CEREBRAL

Se han inventado diversas pruebas clínicas y técnicas de imagenología del cerebro para asistir en el diagnóstico (Cohen y Bookheimer, 1994; Kandel, Schwartz y Jessell, 1995; Raichle, 1994). Debido a que la evaluación neuropsicológica con frecuencia va de la mano con estos procedimientos médicos, se proporciona una breve introducción acerca de las pruebas clínicas y técnicas de imagenología cerebral de más utilización (cuadro 9–6). Son los neurólogos y otros médicos especializados en enfermedades del sistema nervioso quienes autorizan y utilizan exclusivamente estos procedimientos. Además de un examen de consulta que se enfoque en los antecedentes del paciente, su estado mental, reflejos motores, funcionamiento sensorial y habilidad motora, por lo común el neurólogo utiliza una o más pruebas clínicas para ayudarse a diagnosticar o descartar una enfermedad neurológica. Dichas pruebas son esenciales para llegar al diagnóstico correcto del padecimiento médico del paciente. No obstante, no reemplazan a las pruebas neuropsicológicas necesarias para señalar las consecuencias funcionales de los padecimientos neurológicos.

APLICACIONES DE LOS HALLAZGOS DE PRUEBAS NEUROPSICOLÓGICAS

Ahora que se han revisado los aspectos esenciales de la estructura y función del cerebro humano, el lector debería comprender mejor el

Cuadro 9–5. Resumen de los hallazgos sobre lateralización cerebral

Sistema funcional	Predominio del hemisferio izquierdo	Predominio del hemisferio derecho
Visión	Procesamiento del campo visual derecho	Procesamiento del campo visual izquierdo
	Reconocimiento de letras, palabras	Reconocimiento de rostros
Audición	Procesamiento del oído derecho	Procesamiento del oído izquierdo
	Procesamiento de sonidos relacionados con el lenguaje	Procesamiento de música y sonidos ambientales
Somatosensorial	Entrada sensorial del lado derecho	Entrada sensorial del lado izquierdo
Movimiento	Salida motora hacia el lado derecho	Salida motora hacia el lado izquierdo
	Movimiento voluntario complejo, incluyendo habla	
Lenguaje	Habla, lectura, escritura y aritmética	Entonación y patrón emocional del habla
Memoria	Memoria verbal	Memoria pictórica
Procesos espaciales		Análisis de espacio geométrico y visual
Emoción		Comprensión y expresión de emoción
Olfato	Olfato en la fosa nasal izquierda	Olfato en la fosa nasal izquierda

Cuadro 9–6. Pruebas clínicas y técnicas de imagenología del cerebro útiles en el diagnóstico neurológico

Electroencefalografía (EEG)
El EEG produce un registro de la actividad eléctrica de la corteza a partir de electrodos colocados en áreas específicas del cráneo. Las fluctuaciones en actividad se representan como líneas de tinta separadas que se dibujan en un rollo de papel continuo. El EEG es un índice rudimentario, dado que las fluctuaciones de un momento a otro de cada línea de tinta reflejan la actividad sincronizada de millones de neuronas. La prueba es útil para el diagnóstico de trastornos convulsivos y en la localización de actividad cerebral anormal, como la provocada por un tumor.

Angiografía cerebral
En esta técnica se inyecta un tinte especial opaco a las ondas de radio en una de las principales arterias que irrigan al cerebro (arteria vertebral o carótida) y después se toma una radiografía del cerebro. Debido a que el tinte bloquea los rayos X, el sistema arterial del cerebro resalta en evidente relieve en negativo. Así, el médico puede localizar las anormalidades vasculares como un aneurisma (un abultamiento peligroso de una arteria). También, si una arteria se encuentra desplazada de su localización normal, el especialista puede inferir una patología subyacente como la existencia de un tumor. La angiografía tradicional representa un ligero riesgo para el paciente, porque la inyección de tinte puede causar complicaciones neurológicas. Con los avances continuos en la técnica, es probable que la angiografía con resonancia magnética reemplace a los enfoques tradicionales.

Tomografía computarizada (TC)
En un rastreo de TC, un estrecho haz de rayos X atraviesa el cerebro desde docenas de ángulos diferentes; la máquina detecta la cantidad de rayos X que surgen del otro lado. La densidad de las diferentes estructuras internas aparece en películas de radiografía en proporción inversa a su absorción de rayos X. Una computadora resuelve los aspectos matemáticos para reconstruir una representación tridimensional de las densidades internas del cerebro, con lo cual se revelan estructuras importantes. La computadora imprime aproximadamente ocho radiografías transversales bidimensionales del cerebro, cada una desde un plano diferente. La TC produce una resolución de menos de 1 mm; se pueden observar con facilidad los tumores, coágulos de sangre y desplazamientos ventriculares. La TC es menos dañina que una radiografía tradicional de tórax.

Tomografía por emisión de positrones (TEP)
En un rastreo TEP, al paciente se le inyecta una forma radioactivamente marcada de glucosa, el combustible metabólico esencial del cerebro. El nivel de radioactividad es extremadamente limitado y no se considera dañino. Después se registra la radioactividad a través de un detector especial que rodea al cráneo del paciente. Debido a que la glucosa va hacia las partes más activas del cerebro, el rastreo TEP mide el nivel de actividad y no la estructura en sí. Por tanto, se puede utilizar un rastreo TEP para estimar la actividad cerebral regional, lo cual puede ser útil en el diagnóstico de la enfermedad de Alzheimer, esquizofrenia y otros padecimientos en los que existe deterioro cerebral. Los rastreos TEP también pueden utilizarse para localizar los sitios receptores al hacer que el paciente inhale un gas radioactivo que se fija a sitios receptores específicos, como los de dopamina en los ganglios basales. El principal inconveniente del TEP es el nivel de tecnología requerido. Algunas aplicaciones requieren un ciclotrón cercano para la creación de isótopos de vida corta.

Imágenes por Resonancia Magnética (IRM)
El principio funcional de las IRM es que ciertos átomos como el hidrógeno se comportan como pequeños imanes giratorios. Cuando se les coloca en un fuerte campo magnético, estos átomos se alinearán entre sí. Cuando después se lancen ondas de radio a los átomos en ángulos rectos con respecto al campo magnético, los átomos oscilarán de manera sincrónica entre sí. En el momento en que se apagan las ondas de radio, un serpentín de alambre que rodea al cráneo detectará el voltaje o resonancia magnética en el campo magnético, donde el voltaje será más fuerte en las áreas que contienen mayores concentraciones de hidrógeno. Debido a que muchas partes del cerebro son ricas en hidrógeno (en especial aquellas que contienen agua o H_2O), el patrón diferencial de resonancia magnética ayuda a revelar las estructuras cerebrales subyacentes. La resolución espacial de las IRM es tan nítida que las imágenes se asemejan a material anatómico seccionado y fijo. Un nuevo procedimiento que se conoce como angiografía por resonancia magnética, promete reemplazar a procedimientos más peligrosos e invasores como la angiografía cerebral tradicional.

significado neurológico de los resultados en una prueba psicológica; pero en la evaluación neuropsicológica participa más que la mera comprensión de la función cerebral. Se concuerda con Kolb y Whishaw (1990) en que el valor de la neuropsicología no reside sólo en su contribución a una comprensión de la función del cerebro, sino también en la posibilidad de aplicación de sus conocimientos básicos a los problemas humanos. Para ilustrar estas cuestiones entrelazadas —las pruebas con base neuropsicológica conducen a una comprensión y ésta a la aplicación— se regresa al caso del estudiante reprobado en la escuela de preparación para estudios de medicina.

El lector podría desear examinar los resultados de prueba de este estudiante reprobado que se presentan al principio del capítulo (ejemplo de caso 9–1) bajo la perspectiva del conocimiento obtenido de las páginas anteriores. Con propósitos de una revisión rápida, el estudiante había reprobado cursos como química orgánica y embriología a pesar su CI Total superior de 122. Además de su puntuación sospechosamente baja en una prueba de pensamiento abstracto (Prueba de categorías), el estudiante no podía copiar de manera precisa una cruz griega y su velocidad de golpeteo con los dedos para la mano izquierda era bastante lenta en términos comparativos. El lector puede reconocer el primer síntoma —la dificultad para copiar una forma— como dispraxia de construcción, lo cual indica con frecuencia un daño en el hemisferio derecho. Debido a que la dispraxia señala una grave debilidad para manejar relaciones espaciales, la dificultad de este estudiante con la química orgánica y la embriología no sorprende. Además, el segundo síntoma —lentificación motora en la mano izquierda— también sugiere alteración en el hemisferio derecho. La mano no dominante debería ser aproximadamente 10% más lenta que la mano dominante (Reitan y Wolfson, 1993). En este caso, la mano no dominante es 25% más lenta que la dominante (56 contra 42). Aunque la causa continúa siendo un misterio, una tomografía computarizada (TC) confirmó que el estudiante de premedicina había sufrido una lesión estática en la región frontoparietal del hemisferio derecho. Por supuesto, este hecho podría haberse revelado sólo con el uso de la TC, sin referencia a los resultados de una prueba neuropsicológica; sin embargo, la batería de pruebas satisfizo un propósito útil al documentar las consecuencias funcionales del daño cerebral. De manera incidental, el estudiante cambió a una carrera en historia —una búsqueda académica más compatible con sus fortalezas en el hemisferio izquierdo— y se graduó con un título en educación.

RESUMEN

1. La neuropsicología es el estudio de la relación entre función cerebral y conducta. En la evaluación neuropsicológica se utilizan pruebas sensibles a la disfunción cerebral para la evaluación individual.
2. El Sistema Nervioso Central consta de cerebro y médula espinal. El Sistema Nervioso Periférico incluye a los pares craneales y a la red de nervios que emanan de la médula espinal. El cerebro está formado de 10^{11} neuronas y de un número todavía más grande de células gliales que proporcionan apoyo estructural.
3. La parte más baja del cerebro es el metencéfalo, que contiene al bulbo raquídeo. Las fibras nerviosas de las partes superiores del cerebro atraviesan este punto y cruzan en su camino a la médula espinal. La médula ósea ayuda a controlar deglución, vómito, respiración, presión sanguínea, inhalación y frecuencia cardiaca.
4. La formación reticular atraviesa a la médula ósea. Núcleos específicos dentro de la formación reticular se proyectan a amplias áreas del cerebro y de ese modo ayudan a mediar reflejos posturales complejos y tono

muscular. Se sabe que partes de esta estructura conocidas como sistema reticular activador gobiernan la excitación general o consciencia.

5. Localizado dentro del mesencéfalo, el tegmento es una estación de relevo para fibras sensoriales y motoras, y también contiene muchos de los pares craneales. Los 12 pares craneales ayudan a gobernar los principales sentidos (olfato, visión, gusto, oído) y también participan en el movimiento de rostro y lengua.
6. Las estructuras del diencéfalo incluyen al tálamo, hipotálamo y glándula pineal. La función de esta última no es clara, aunque se sabe que ayuda a regular los ritmos biológicos cíclicos a través de la liberación de la hormona melatonina.
7. El tálamo es una estructura importante que proporciona información y datos sensoriales a la corteza cerebral acerca del movimiento continuo; también puede participar en la memoria, atención, habla y experiencia emocional, ya sea de manera directa o indirecta a través de su función como estación de relevo para sitios cerebrales distantes.
8. El hipotálamo participa en numerosos aspectos de la conducta motivada, incluyendo alimentación, conducta sexual, sueño, regulación de temperatura, conducta emocional y movimiento; también controla a la glándula hipofisaria, con lo cual modula una amplia variedad de funciones endocrinas.
9. La formación más exterior del prosencéfalo es el telencéfalo, que incluye al lóbulo límbico, ganglios basales, cuerpo calloso y corteza cerebral. El lóbulo límbico, incluyendo al hipocampo, tabique y circunvolución cingulada, participa en la regulación de la emoción. El hipocampo tiene una función importante en la consolidación de experiencias nuevas en recuerdos a largo plazo.
10. Los ganglios basales, que incluyen al núcleo caudado, putamen y globo pálido, ayudan a gobernar el movimiento. El mal de Parkinson es un trastorno degenerativo de los ganglios basales que se caracteriza por movimientos involuntarios, incluyendo temblor; pobreza y lentitud de movimiento, sin parálisis; y cambios en postura y tono muscular.
11. El cuerpo calloso es la comisura principal que sirve para integrar las funciones de ambos hemisferios. Mucho se ha aprendido acerca de la función cerebral a través del estudio de personas con epilepsia, a quienes se ha seccionado quirúrgicamente el cuerpo calloso por propósitos terapéuticos.
12. La capa más exterior del cerebro es la corteza cerebral o neocorteza. La corteza cerebral comprometida consta de sitios dedicados al procesamiento sensorial básico de la visión, audición, tacto y control motor; pero su mayor parte es corteza no comprometida o de asociación, y se utiliza en el análisis de información sensorial y en la formulación de respuestas motoras.
13. Los lóbulos frontales participan en la programación y regulación de la planificación de metas, incluyendo al movimiento corporal; los lóbulos parietales median la conciencia corporal contralateral y ayudan con tareas espaciales. Los lóbulos occipitales participan en la visión, y los temporales contribuyen al procesamiento auditivo, memoria e impulsos biológicos como agresión, temor y sexualidad.
14. En la mayoría de las personas, el hemisferio cerebral izquierdo es dominante para el lenguaje. El área de Broca en el lóbulo frontal está al servicio de los componentes expresivos del lenguaje, mientras los componentes receptivos se median a través del área de Wernicke en el lóbulo temporal.
15. Las funciones especializadas del hemisferio cerebral derecho incluyen el análisis del espacio geométrico y visual, la comprensión y expresión de emociones, el procesamiento de la música y de sonidos ambientales no verbales, y el reconocimiento táctil de formas complejas.

16. Se pueden utilizar varias pruebas clínicas y técnicas de imagenología del cerebro para ayudar a diagnosticar los padecimientos neurológicos. Estos métodos incluyen al EEG, arteriografía cerebral, tomografía computarizada, tomografía por emisión de positrones, imagenología por resonancia magnética y angiografía por resonancia magnética.

TÉRMINOS Y CONCEPTOS CLAVE

Neuropsicología	p. 364	Lóbulo límbico	p. 370
Ventrículos	p. 366	Hipocampo	p. 370
Metencéfalo	p. 366	Ganglios basales	p. 371
Médula ósea	p. 366	Mal de Parkinson	p. 371
Formación reticular	p. 367	Cuerpo calloso	p. 371
Cerebelo	p. 367	Corteza cerebral	p. 372
Mesencéfalo	p. 368	Lóbulos occipitales	p. 375
Pares craneales	p. 369	Agnosia visual	p. 375
Prosencéfalo	p. 369	Lóbulos parietales	p. 376
Glándula pineal	p. 369	Lóbulos temporales	p. 377
Tálamo	p. 369	Lóbulos frontales	p. 378
Hipotálamo	p. 370	Dispraxia de construcción	p. 385

TEMA 9B
Evaluación neuropsicológica y geriátrica

Un modelo conceptual de las relaciones cerebro-conducta
Evaluación de la entrada sensorial
Medidas de atención y concentración
Pruebas de aprendizaje y memoria
Evaluación de las funciones del lenguaje
Pruebas de capacidad espacial y de manipulación
Evaluación de las funciones ejecutivas
Evaluación de la salida motora
Baterías de prueba en evaluación neuropsicológica
Ejemplo de caso 9–2. Diagnóstico neuropsicológico Luria-Nebraska (LNNB)
Evaluación del estado mental en ancianos
Resumen
Términos y conceptos clave

Las pruebas y procedimientos neuropsicológicos engloban una variedad ecléctica de métodos y propósitos. En un extremo del espectro se encuentran las pruebas sencillas, de detección, con duración de 10 minutos, que se utilizan para sondear la necesidad de una mayor evaluación. En el otro extremo están las baterías exhaustivas, de seis horas, diseñadas para proporcionar una evaluación completa. Entre estos dos puntos se encuentran miles de instrumentos especializados que se han desarrollado para medir capacidades neuropsicológicas particulares, los cuales a primera vista parecerían resistir la categorización simple, como si los investigadores en esta área hubieran seguido una filosofía incoherente, al tanteo, en el desarrollo de nuevos instrumentos y procedimientos. No obstante, en un escrutinio más completo se hace evidente que la mayor parte de las pruebas neuropsicológicas encajan dentro de un modelo simple y lógico de relaciones cerebro-conducta, que se utilizará como esquema para el análisis de las pruebas y procedimientos neuropsicológicos más conocidos.

La evaluación neuropsicológica implica más que la aplicación y calificación de pruebas especializadas. Un componente importante de cualquier evaluación es la valoración del estado

mental del cliente. Esto es particularmente cierto en el caso de clientes ancianos que pueden presentar enfermedad de Alzheimer u otras formas de demencia. De acuerdo con esto, se cierra el capítulo enfatizando la evaluación del estado mental en los ancianos.

UN MODELO CONCEPTUAL DE LAS RELACIONES CEREBRO-CONDUCTA

Bennett (1988) ha propuesto un modelo simplificado de relaciones cerebro-conducta útil para organizar la profusión aparentemente caótica de pruebas neuropsicológicas (figura 9–6). Su concepto es una pequeña expansión del modelo presentado por Reitan y Wolfson (1993). Según este punto de vista, cada prueba o procedimiento neuropsicológico valora una o más de las siguientes categorías:

1. Entrada sensorial
2. Atención y concentración
3. Aprendizaje y memoria
4. Lenguaje
5. Capacidad espacial y de manipulación
6. Funciones ejecutivas:
 - Análisis lógico
 - Formación de conceptos
 - Razonamiento
 - Planificación
 - Flexibilidad de pensamiento
7. Salida motora

El orden de las categorías listadas corresponde aproximadamente al orden en que el cerebro analiza la información que ingresa para una respuesta o salida motora.

En lo que resta de este tema, el examen de las pruebas y procedimientos neuropsicológicos se organiza alrededor de estas siete categorías. Dentro de cada una, se revisarán las pruebas establecidas y se presentarán nuevos instrumentos que prometen ampliar los horizontes

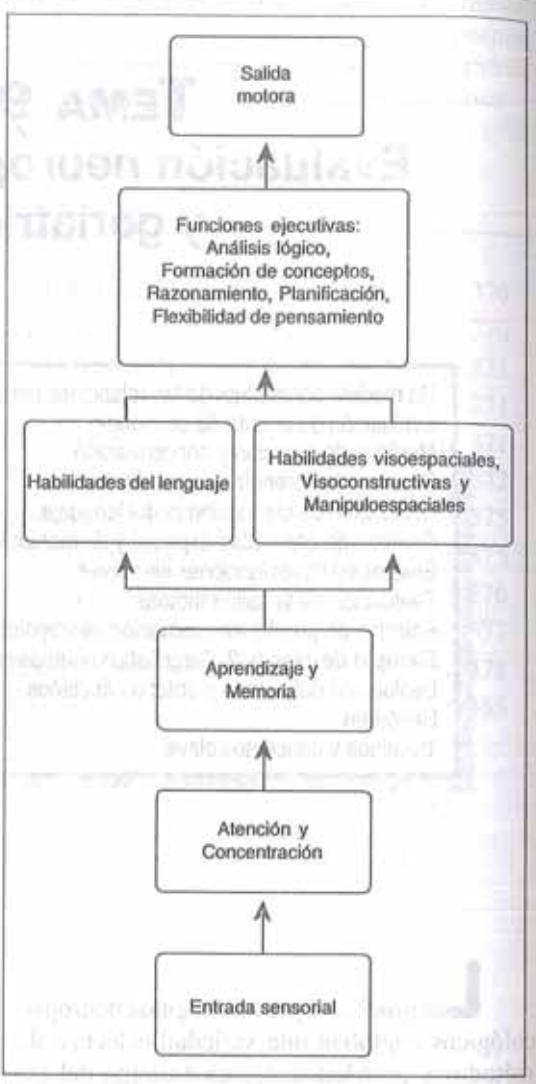


Figura 9–6. Modelo conceptual de las relaciones cerebro-conducta.
Fuente: Basado en Reitan y Wolfson (1993) y reproducido con autorización de Bennett, T. (1988). Use of the Halstead-Reitan Neuropsychological Test Battery in the assessment of head injury. *Cognitive Rehabilitation 6*, 18-25.

de la evaluación neuropsicológica. Sin embargo, el lector necesita conocer que, por lo general, la evaluación neuropsicológica implica una batería de pruebas. Un enfoque es el examen flexible o centrado en el paciente en el que se diseña una batería individualizada de pruebas para cada uno. Estas baterías se basan en la

dolencia exhibida, en cuestiones de canalización y en una evaluación inicial (Goodglass, 1986; Kane, 1991). No obstante, es más común que los neuropsicólogos empleen una batería fija de pruebas para la mayoría de las personas canalizadas; en el cuadro 9–7 se describe una de las más usadas, la *Halstead-Reitan Neuropsychological Battery* (Batería Neuropsicológica Halstead-Reitan). El capítulo se cierra con otra conocida batería fija, la *Luria-Nebraska Neuropsychological Battery* (Diagnóstico Neuropsicológico Luria-Nebraska).

EVALUACIÓN DE LA ENTRADA SENSORIAL

La precisión de la entrada sensorial es fundamental para la suficiencia de la percepción, pensamiento, planes y acción. Un individuo que no vea los estímulos de manera correcta, escuche los sonidos con precisión o procese el tacto de modo confiable, puede enfrentar impedimentos adicionales en los niveles más altos de la percepción y cognición. La evaluación neuropsicológica siempre incorpora un examen multimodal de las capacidades sensoriales.

EXAMEN SENSOPERCEPTUAL

Los procedimientos que desarrollaron Reitan y Klove son típicos de los procedimientos sensoperceptuales (Reitan, 1984, 1985). El *Reitan-Klove Sensory-Perceptual Examination* (Examen Sensoperceptual Reitan-Klove) consta de varios métodos para presentar estímulos unilaterales y bilaterales en las modalidades de tacto, oído y visión. Las tareas son tan sencillas que las personas normales pocas veces cometen algún error. Por ejemplo, se le pide a la persona evaluada que diga cuál mano se le ha tocado (con los ojos cerrados), o que informe cuál oído ha recibido un chasquido de dedos apenas audible o que identifique cuál número se ha delineado sobre la yema del dedo. Los resultados de esta prueba son particularmente diagnósticos si la persona comete más errores de un lado del cuerpo que del otro. El lector recordará del capítulo anterior que la inervación neuronal es casi exclusivamente contralateral. Lo que es más, ciertas áreas de la corteza cerebral se dedican al procesamiento primario del tacto, oído y visión. Así, una persona que tenga dificultad para procesar el tacto en la mano derecha, puede tener una lesión en la circunvolución poscentral del lóbulo parietal izquierdo. De manera similar, la dificultad para procesar el sonido en el oído derecho, puede indicar una lesión en la parte superior del lóbulo temporal izquierdo, y los defectos visuales en el lado derecho pueden indicar daño cerebral en el lóbulo occipital izquierdo.

PRUEBA DE LOCALIZACIÓN DE DEDOS

La *Finger Localization Test* (Prueba de Localización de Dedos) es un procedimiento que desarrollaron los neurólogos para valorar posibles pérdidas sensoriales producidas por deterioro de las funciones cerebrales. La mayor parte de las baterías de pruebas neuropsicológicas emplean una variante de esta prueba, en la que las personas evaluadas deben identificar los dedos que se les han tocado (sin ver). Benton ha desarrollado una prueba bien normada de 60 reactivos de localización de dedos que consiste en tres partes: 1) con la mano visible se identifican dedos individuales que ha tocado el examinador con la punta de un lápiz (10 ensayos con cada mano); 2) con la mano oculta de la vista, se identifican los dedos individuales que ha tocado el examinador (10 ensayos con cada mano); 3) con la mano oculta de la vista, se identifican pares de dedos tocados de manera simultánea por el examinador (10 ensayos en cada mano). El método de respuesta se deja al paciente; nombrar, tocar o señalar los dedos en un diagrama (Benton, Sivan, Hamsher, Varney y Spreen, 1994). Cada presentación de un estímulo se califica como correcto o incorrecto y, por lo común, los adultos sanos cometen muy pocos errores en los 60 ensayos. Las puntuaciones medias para estos adultos son casi

Cuadro 9–7. Batería de pruebas Halstead-Reitan
(Tests and Procedures of the Halstead-Reitan Test Battery)

Prueba	Descripción
Prueba de categorías:*	Mide razonamiento abstracto y formación de conceptos; requiere que la persona evaluada encuentre la regla para categorizar dibujos de formas geométricas.
Prueba de desempeño táctil:*	Mide capacidad cinestésica y sensoriomotora; requiere que la persona con los ojos vendados coloque los cubos en un orificio con el corte adecuado en un tablero vertical con la mano dominante, luego con la no dominante y después con ambas manos; también evalúa la memoria incidental de los cubos.
Prueba de percepción de sonidos del lenguaje:*	Mide atención y síntesis audiovisual; requiere que la persona elija entre cuatro opciones la versión escrita de palabras sin sentido mostradas en cinta de audio.
Prueba rítmica Seashore:*	Mide atención y percepción auditiva; requiere que la persona evaluada indique si un par de ritmos musicales son iguales o diferentes.
Prueba de golpeteo con los dedos:*	Mide velocidad motora; requiere que el individuo golpetee una palanca parecida a la tecla de un telégrafo con la mayor rapidez posible durante 10 segundos.
Fortaleza de asimiento:	Mide fortaleza de asimiento con un dinamómetro; requiere que la persona evaluada apriete con la mayor fuerza posible; hay ensayos separados con cada mano.
Trazado de un camino, partes A, B:	Mide capacidad de rastreo, flexibilidad mental y velocidad; requiere que la persona conecte número (parte A) o números y letras en orden alternante (parte B) con una línea de lápiz, bajo presión de tiempo.
Reconocimiento táctil de formas:	Mide capacidad sensoperceptual; requiere que la persona evaluada reconozca formas sencillas (p. ej., triángulo) colocado en la palma de la mano.
Examen sensoperceptual:	Mide capacidad sensoperceptual; requiere que la persona responda a tareas sensoriales bilaterales simples; por ejemplo, detectar cuál dedo se le ha tocado, cuál oído ha recibido un sonido breve; evalúa los campos visuales.
Prueba de detección de la afasia:	Mide capacidades expresivas y receptivas del lenguaje; las tareas incluyen nombrar un artículo representado en una ilustración (p. ej., tenedor), repetición de frases cortas; se incluyen tareas de copiado (no es una medida de la afasia) debido a razones históricas.
Complementaria	WAIS-III, WRAT-3, MMPI-2, pruebas de memoria como la Escala Wechsler de Memoria–III o la *Rey Auditory Verbal Learning Test.*

* En términos estrictos, estas cinco medidas constituyen a la Batería de pruebas Halstead-Reitan. No obstante, en un sentido de referencia cotidiana, la Halstead-Reitan incluye todas las medidas listadas antes.

perfectas, y van desde 56 hasta 60 en varias muestras. En contraste, aquéllos con padecimientos cerebrales enfrentan un desafío en la localización de dedos, en particular en la segunda y tercera partes de la prueba.

MEDIDAS DE ATENCIÓN Y CONCENTRACIÓN

La capacidad de atención del cerebro hace posible atender a estímulos significativos, separar la entrada sensorial irrelevante de la profusión de estímulos que ingresan y cambiar a estímulos alternativos cuando las condiciones lo demandan (Kinsbourne, 1994). Aunque en teoría sería posible hacer distinciones sutiles entre atención simple, concentración, viraje mental, rastreo mental, vigilancia y otras variedades de atención/concentración, en la práctica es difícil separar estas habilidades. Sólo una medida de atención —la *Test of Everyday Attention*, TEA (Prueba de Atención Cotidiana)— ha tenido éxito en dividir la atención en sus fuentes componentes. En las siguientes secciones se analiza la TEA y otras notables medidas de deterioro en la atención.

PRUEBA DE ATENCIÓN COTIDIANA

La *Test of Everyday Attention*, TEA (Prueba de Atención Cotidiana), es una medida prometedora que diseñaron en Gran Bretaña Robertson, Ward, Ridgeway y NimmoSmith (1994, 1996); mide los subcomponentes de la atención, incluyendo atención sostenida, selectiva, dividida y cambio de atención. En el cuadro 9–8 se describen las subpruebas de la TEA. La prueba tiene tres versiones paralelas que se han validado de manera adecuada con clientes que han sufrido lesiones cefálicas cerradas, pacientes con accidentes cerebrovasculares y personas con enfermedad de Alzheimer. Los datos normativos se basan en el desempeño de 154 individuos sanos entre los 18 y 80 años de edad. Las personas evaluadas disfrutan de los escenarios de la vida real de la TEA, lo cual se añade a la validez ecológica del instrumento. Esta prueba es sumamente sensible a los efectos normales debidos a la edad en la población general y, por tanto, es muy adecuada para la evaluación geriátrica. Con excepción de la subprueba de Cuenta de Ascensor, las ocho subpruebas se estandarizaron para producir puntuaciones equivalentes con una media común de 10 y desviación estándar de 3. Así, la TEA permite el análisis de subprueba como medio para identificar las fortalezas y debilidades particulares de un individuo (Crawford, Sommerville y Robertson, 1997).

Cuadro 9–8. Subpruebas de la Prueba de atención cotidiana (TEA)

Búsqueda de mapa: una búsqueda cronometrada de dos minutos de 80 símbolos en un mapa de colores; mide atención selectiva.

Cuenta de ascensor: simulación de una cuenta de los pisos en el ascensor con tonos presentados en cinta de audio; mide atención sostenida.

Cuenta de ascensor con distracción: igual que la anterior, pero con distracciones auditivas; mide atención sostenida.

Ascensor visual: simulación visual de la cuenta de los pisos en el ascensor con inversiones hacia arriba-hacia abajo; mide cambio de atención.

Ascensor auditivo con inversión: igual que el ascensor visual, excepto que se presenta en cinta de audio; mide cambio de atención.

Búsqueda telefónica: búsqueda de símbolos clave al mismo tiempo que se buscan los registros en un directorio telefónico simulado de anuncios clasificados; mide atención dividida.

Tarea dual de búsqueda telefónica: combina la Búsqueda telefónica con la cuenta simultánea de tonos auditivos; mide atención dividida.

Lotería: el individuo escucha los números ganadores en busca de aquellos que terminan en 55 y después escribe los estímulos anteriores; mide atención sostenida.

PRUEBA DE DESEMPEÑO CONTINUO

La *Continuous Performance Test*, CPT (Prueba de Desempeño Continuo), no es en realidad

una sola prueba, sino más bien una familia de procedimientos similares que proviene de la investigación líder de Rosvold, Mirsky, Sarason y colaboradores (1956). Estos autores diseñaron una medida de atención sostenida (llamada también vigilancia) que implicaba la presentación continua de letras sobre una pantalla. En algunos casos, las personas evaluadas tenían que presionar una tecla cuando aparecían ciertas letras (p. ej., *x*); en otros, los individuos tenían que presionar una tecla cuando cierta letra aparecía **después** de otra (p. ej., la *x* cuando ocurría después de la *a*). Los errores de omisión son aquéllos en los que la persona no presiona la tecla al presentarse el estímulo blanco. Los errores de comisión son aquellos que realiza el individuo cuando presiona la tecla al presentarse un estímulo no blanco. Las personas sanas cometen pocos errores. El método ha resultado especialmente eficaz en la identificación de trastornos por daño cerebral que incluyen hiperactividad, efectos de las drogas, esquizofrenia y daño cerebral explícito (Huhtaniemi, Haier, Fedio y Buchsbaum, 1983; Sostek, Buchsbaum y Rapoport, 1980).

La CPT es ideal para la adaptación a computadora y en la literatura han aparecido docenas de diferentes versiones de ella (p. ej., Buchsbaum y Sostek, 1980; Gordon y Mettelman, 1988). Por desgracia, la proliferación de pruebas similares, pero no idénticas, ha obstaculizado la investigación acerca de la utilidad práctica de esta prometedora medida de la atención. Recientemente, Sandford y Turner (1997) publicaron una CPT por computadora que emplea tanto estímulos visuales como auditivos. Las normas de la *Intermediate Visual and Auditory Continuous Performance Test*, IVA (Prueba Visual y Auditiva Intermedia del Desempeño Continuo), se establecieron con 781 personas normales cuyas edades iban de 5 a 90 años y evaluadas para detectar déficit de atención, dificultades de aprendizaje, problemas emocionales y uso de medicamentos. En un análisis, la IVA mostró 92% de sensibilidad (es decir, una tasa de 8% de falsos negativos) y 90% de especificidad (es decir una tasa de 10% de falsos positivos) en cuanto a diferenciar a niños con un diagnóstico de Trastorno por Déficit de Atención con Hiperactividad (TDAH) de niños sanos. Este instrumento es sólo una de las muchas pruebas neuropsicológicas prometedoras que toma ventaja de la tecnología de las microcomputadoras.

TAREA RÍTMICA AUDITIVA DE SUMA EN SERIE

Si se considera su absoluta sencillez, la *Paced Auditory Serial Addition Task*, PASAT (Tarea Rítmica Auditiva de Suma en Serie), es un índice extremadamente sensible del rastreo mental (Gronwall y Wrightson, 1974; Gronwall y Sampson, 1974). La persona evaluada escucha una serie de dígitos que se presentan en cinta de audio y suma cada par sucesivo de dígitos. Así, si los números que se presentan son "3-1-9-5-4", el individuo debería responder "4-10-14-9".

La PASAT comienza con una serie de práctica de 10 dígitos, donde cada número nuevo se presenta cada 2.4 segundos. La prueba en sí consiste en 61 estímulos (lo cual requiere 60 sumas) en cada una de cuatro velocidades de presentación: 2.4, 2.0, 1.6 y 1.2 segundos entre dígitos. A través del cálculo del porcentaje correcto en cada una de las tasas de presentación, el examinador obtiene cuatro puntuaciones en la PASAT.

Aunque el formato de la PASAT es simple, las demandas de procesamiento de información de esta prueba son un tanto arduas. A fin de tener un desempeño adecuado, la persona examinada debe conservar dos números en la memoria a corto plazo, llevar a cabo una suma mental, decir la respuesta, retener en la memoria a corto plazo sólo el último de los dos números, anexar el último dígito a la memoria a corto plazo y después comenzar el ciclo de nuevo. Las personas con deterioro en las funciones cerebrales encontrarán que este malabarismo mental es cognitivamente abrumador.

Gronwall recomienda la PASAT para la evaluación en serie de pacientes con conmoción cerebral (Gronwall, 1977; Gronwall y

Wrightson, 1981). Definida de manera breve, la **conmoción cerebral** es una alteración transitoria de la conciencia debido a un golpe en la cabeza, a la que le puede seguir amnesia temporal, mareo, náuseas, pulso débil y respiración lenta, no obstante, no existe un daño cerebral orgánico demostrable (McMordie, 1988). La PASAT tiene amplio reconocimiento como un instrumento muy sensible a los efectos de la conmoción cerebral (Stuss, Stethem, Hugenholtz y Richard, 1989). Sin embargo, una cuestión crucial de ésta es cuánto debería tardar el paciente en recuperarse. Cuando las puntuaciones sucesivas en la PASAT regresan finalmente al rango normal —lo cual podría tomar varios días o semanas— el terapeuta puede tener mayor confianza en que el paciente está listo para regresar al trabajo.

En gran medida debido a una fundamentación humanista, Lezak (1995) advierte en contra del uso rutinario de la PASAT. En su experiencia, incluso las personas totalmente cognitivas intactas sienten que la prueba es muy estresante y piensan que sus respuestas son incorrectas, aunque su desempeño sea normal. Dado que existen modos más fáciles de demostrar el deterioro en la atención, esta autora recomienda el uso de la PASAT sólo en circunstancias especiales:

> ... La tengo disponible para esas ocasiones en que es necesario hacer obvias las deficiencias sutiles en atención a los escépticos más retrógrados, con algún propósito que en mucho está en beneficio del paciente; y entonces preparo de antemano a estos pacientes, haciéndoles saber que puede ser un procedimiento desagradable y que pueden sentir que van fallando aunque no sea así.

En un campo donde las pruebas se justifican principalmente con base en las tasas de acierto y cosas similares, su perspectiva centrada en el paciente es buena.

RESTA DE SERIES DE SIETE

Una tarea bien conocida que se incluye con frecuencia en el examen del estado mental es la *Subtracting Serial Sevens* (Resta de Series de Siete) (Strub y Black, 1985). Se le pide al examinando que "reste siete de 100", cuando termina se le dice "ahora reste 7 de 93 y siga restando sietes hasta que ya no pueda continuar más". El examinador registra el número de restas individuales incorrectas y también puede dar puntuaciones al tiempo que se requiere y las pausas mayores a cinco segundos.

Smith (1967) ha informado sobre uno de los estudios normativos sobre la Resta de Series de Siete. Este autor evaluó a 132 empleados adultos, la mayoría con títulos universitarios o profesionales. Sólo 2% de su muestra no pudo terminar la prueba, mientras otro 5% cometió más de cinco errores. Las mujeres estaban más propensas a los errores que los varones, en particular mujeres mayores de 45 años, sin educación universitaria. Así, los examinadores no deben sobreinterpretar los problemas menores con la Resta de Series de Siete. Por otro lado, el desempeño gravemente defectuoso —una incapacidad para proseguir, tasa muy elevada de errores o restas muy lentas— es característico de individuos con daño cerebral.

MEDIDAS ADICIONALES DEL DETERIORO EN LA ATENCIÓN

Muchas de las pruebas diseñadas para otros propósitos poseen un fuerte factor asociado con la atención. Las subpruebas de Retención de Dígitos y Aritmética de las escalas Wechsler de inteligencia tienen reconocimiento como buenos índices de atención auditiva inmediata. Las subpruebas de Claves o Símbolos y Dígitos también tienen fuertes cargas factoriales en el factor de libertad de la distractibilidad.

Smith (1968, 1973) ha diseñado una interesante extensión de la subprueba de Dígitos y Símbolos de Wechsler, conocida como *Symbol Digit Modalities Test*, SDMT (Prueba de Modalidades de Símbolos y Dígitos). En esta prueba, los símbolos están impresos en una página y la persona evaluada escribe debajo de ellos los números correspondientes (figura 9–7). Con este formato, el examinador puede aplicar

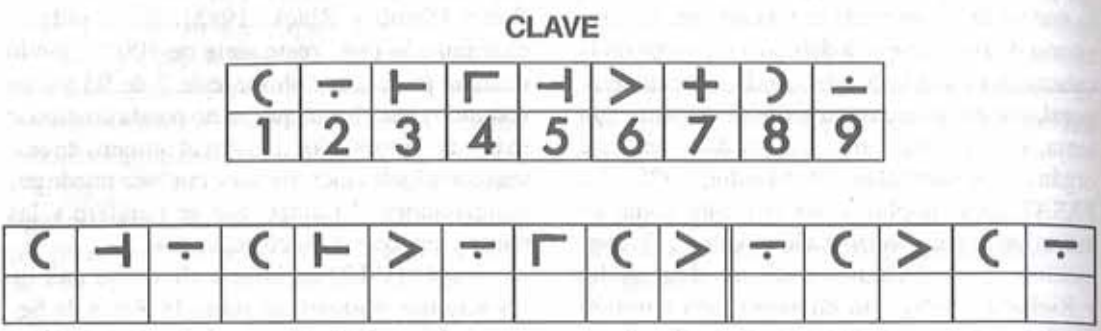


Figura 9–7. Prueba de modalidades de símbolos y dígitos.
Fuente: Reproducido con autorización de Smith, A. (1973). *Symbol Digit Modalities Test Manual.* Los Angeles: Western Psychological Services. *Copyright* del material de la Symbol Digit Modalities Test © 1973 de Western Psychological Services. Reproducido con autorización del editor, Western Psychological Services, 12031 Wilshire Boulevard, Los Angeles, California 90025, Estados Unidos de Norteamérica.

tanto un ensayo escrito como uno oral, lo cual ayuda a aislar la fuente de dificultad con la tarea de sustitución. Por ejemplo, una persona que presenta un deterioro normativo en la parte escrita, pero cuyo desempeño es superior al promedio en la parte oral, podría sufrir deterioro en el control motor. La correlación entre las puntuaciones de la SDMT y de Dígitos y Símbolos es sumamente elevada ($r = .91$), pero la SDMT produce puntuaciones que son comparativamente inferiores a las de Dígitos y Símbolos (Morgan y Wheelock, 1992).

Varias pruebas de la batería Halstead-Reitan son buenas medidas de atención (Bennett, 1988). En la *Speech Sounds Perception Test*; SSPT (Prueba de Percepción de Sonidos del Habla), la persona evaluada debe elegir entre cuatro opciones la versión escrita de palabras grabadas que no tiene sentido. Por ejemplo, la voz en la cinta podría decir "frip", mientras que la persona evaluada debe leer las cuatro opciones y subrayar la alternativa correcta:

Frib, flib, frip, flip.

La SSPT es sumamente sensible a los deterioros en la atención que provienen de cualquier tipo de daño cerebral. La *Seashore Rythym Test* (Prueba Rítmica Seashore), que originalmente era una prueba de aptitud musical en la que deben compararse pares de ritmos musicales, también resulta depender en gran medida de los procesos de atención. La *Trail Making Test* (Prueba de Trazado de un Camino), partes A y B, también es sensible al deterioro en la atención. Shum, McFarland y Bain (1990) analizan pruebas adicionales de atención.

PRUEBAS DE APRENDIZAJE Y MEMORIA

El aprendizaje y la memoria son procesos entrelazados difíciles de analizar de manera aislada. En general, el aprendizaje de material novedoso requiere del ejercicio de la memoria. Lo que es más, muchas pruebas de memoria incorporan una curva de aprendizaje a través de aplicaciones repetidas. La separación de los procesos de aprendizaje y memoria es teóricamente posible, pero es de poco valor práctico en la evaluación clínica. En nuestro caso no se realiza una estrecha distinción entre estos procesos.

Las pruebas de memoria pueden clasificarse según varias dimensiones que incluyen corto plazo contra largo plazo, verbal contra pictórica y de curva de aprendizaje contra sin curva de aprendizaje. Estas dimensiones reflejan los factores neurológicos que se examinan en la sección anterior. Por ejemplo, la memoria verbal está significativamente lateralizada hacia el hemisferio izquierdo, mientras la memo-

ria pictórica se sirve en gran medida del hemisferio derecho. El lector interesado puede consultar a Lezak (1995) y Reeves y Wedding (1994) para análisis más detallados de los sustratos neuronales de los diferentes tipos de memoria. Aquí se centrará la atención en las características psicométricas de cuatro pruebas bastante diferentes de la memoria.

ESCALA WECHSLER DE MEMORIA–III

La *Wechsler Memory Scale-III*, WMS-III (Escala Wechsler de Memoria–III; Tulsky, Zhu y Ledbetter, 1997), es una revisión sustancial de una prueba sencilla de una página que se publicó hace más de 50 años (Wechsler, 1945). La tercera edición es una prueba extensa y multifásica de la memoria que consta de 17 subpruebas, incluyendo siete opcionales. Las 10 subpruebas principales se describen en el cuadro 9–9; constituyen la base para las puntuaciones escalares obtenidas con ajuste por edad (media de 100 y DE de 15) para **ocho** índices primarios de memoria:

Auditiva inmediata	Auditiva demorada
Visual inmediata	Visual demorada
Memoria inmediata	Reconocimiento auditivo demorado
Memoria general	Memoria simultánea

Las normas de la WMS-III se establecieron junto con las del WAIS-III en 1997. La estandarización del nuevo instrumento es espléndida, con 200 casos seleccionados para cada una de estas bandas por edad: 16 a 17, 18 a 19, 20 a 24, 25 a 29, 30 a 34, 35 a 44, 45 a 54, 55 a 64, 65 a 69, 70 a 74, 75 a 79. Para las dos bandas de mayor edad (80 a 84, 85 a 89) se incluyeron 150 y 100 casos, respectivamente. Con base en los datos del censo de 1995, los participantes de la muestra de estandarización se estratificaron de manera cuidadosa en cuanto a edad, sexo, raza/origen étnico, nivel educativo y región geográfica.

Los estudios de validez de la WMS-III son sumamente positivos, aunque la investigación

Cuadro 9–9. Escala Wechsler de Memoria–III Subpruebas principales

Subpruebas de recuerdo inmediato

Memoria lógica I: recuerdo de elementos esenciales de historias breves leídas a la persona evaluada.

Rostros I: recuerdo de sí-no de 24 rostros blanco presentados cada uno durante dos segundos.

Pares verbales asociados I: recuerdo de una lista de términos apareados (p. ej., camión-flecha) cuando sólo se presenta el primer término (p. ej., camión–?).

Retratos de familia I: recuerdo de la localización y actividades de personas representadas en ilustraciones de escenas familiares.

Secuencias de letras y números: reordenación de dígitos y letras distribuidas de manera aleatoria de modo que ambos se encuentren en orden correcto (p. ej., "7, x, d, s, 4, 2" se ordenan como "2, 4, 7, d, s, x).

Retención espacial: un análogo visual de Retención de dígitos en el que cubos numerados se colocan en un orden particular; la persona evaluada debe llevar a cabo tanto la serie en sentido progresivo como en sentido inverso.

Subpruebas de recuerdo demorado*

Memoria lógica II

Rostros II

Pares verbales asociados II

Retratos de familia II

* Recuerdo demorado de los estímulos de la primera aplicación después de 30 minutos.

analítico factorial no siempre sustenta el análisis designado en los diversos aspectos de memoria que se citan antes. La evidencia más poderosa de la validez es que el instrumento tiene un funcionamiento adecuado en la detección de deficiencias de memoria. En los estudios iniciales de validación (Tulsky, Zhu y Ledbetter, 1997), se observó que los grupos clínicos con trastornos neurológicos (p. ej., enfermedad de Alzheimer, lesión traumática cefálica) tuvieron puntuaciones significativamente más bajas en los ocho índices primarios de la WMS-III. Por ejemplo, una muestra de 35 individuos con

probables etapas iniciales de la enfermedad de Alzheimer obtuvieron puntuaciones promedio a la mitad o parte superior del rango de 60 puntos en 6 de los 8 índices. Esto es de especial importancia debido a que la deficiencia en la memoria es la afección inicial en la progresión de la enfermedad de Alzheimer.

La investigación de validez con la WMS-III es escasa en este momento. No obstante, la última edición conserva los aspectos esenciales de su antecesora, la WMS-R, para la cual se encuentra disponible gran cantidad de investigación sobre validez. Por ejemplo, Ryan y Lewis (1988) encontraron deficiencias sustanciales de memoria en alcohólicos que habían atravesado recientemente por un proceso de desintoxicación. Éste es un dato importante, ya que los estudios clínicos con la WMS original no revelaron deficiencias de memoria en alcohólicos, lo cual hacía que los investigadores dudaran de la validez de la primera edición. La WMS-R también funciona bien en la identificación de deficiencias neuropsicológicas producidas por lesiones cefálicas cerradas (Reid y Kelly, 1993). En una conclusión relacionada, Mittenberg, Azrin, Millsaps y Heilbronner (1993) encontraron que los individuos que intentan simular los síntomas de lesión cefálica en la WMS-R, producen un patrón de puntuaciones que puede diferenciarse de aquel de personas con un verdadero caso de este tipo de lesión. Ésta es una conclusión importante, debido a que, en general, la precisión de los resultados de prueba se pone en duda cuando las personas con lesión cefálica inician demandas por daños o compensación en el trabajo. La WMS-R también revela las deficiencias esperadas en la memoria en pacientes con esquizofrenia, lo cual sustenta la validez de la prueba (Gold, Randolph, Carpenter y cols., 1992).

PRUEBA REY DE APRENDIZAJE AUDITIVO VERBAL

Esta medida corta y de fácil aplicación de la memoria y aprendizaje verbales apareció inicialmente en francés (Rey, 1964), pero Lezak (1983, 1995) y otros investigadores han proporcionado una adaptación al inglés. Para aplicar la *Rey Auditory Verbal Learning Test*, RAVLT (Prueba Rey de Aprendizaje Auditivo Verbal), el examinador lee una lista de 15 sustantivos concretos a una velocidad de uno por segundo. La persona evaluada recuerda tantos como le sea posible, en cualquier orden. El examinador lee la lista completa en una segunda ocasión, advirtiendo a la persona que debe recordar todas las palabras. Después se procede con una tercera, cuarta y quinta aplicación y rememoración, a las que le sigue un ensayo de interferencia con una nueva lista de palabras. A continuación se evalúa el recuerdo inmediato de la lista original (sin el beneficio de una nueva presentación). Por último, se incluye un ensayo de reconocimiento en el que la persona evaluada debe subrayar las palabras aplicadas en un largo párrafo por escrito. La prueba produce varias puntuaciones, incluyendo el número recordado (de 15) para cada uno de los primeros cinco ensayos, el total de éstos (75 posibles), el recuerdo inmediato después de que se lee la lista de distracción y después la puntuación de reconocimiento.

Rosenberg, Ryan y Prifitera (1984) concluyeron que la RAVLT tiene buen desempeño en la identificación de pacientes de los que se sabe, según otros criterios, que sufren de deterioro en la memoria. Además de una reducción general en el desempeño, los pacientes con deterioro de la memoria mostraron una tasa reducida de mejoría a través de los cinco ensayos. Las normas de la RAVLT para adultos pueden encontrarse en Geffen, Moar, O'Hanlon, Clark y Geffen (1990) y en Wiens, McMinn y Crossen (1988). Las normas para niños de 5 a 16 años de edad se proporcionan en Bishop, Knights y Stoddart (1990). Ivnik, Malec, Smith y colaboradores (1992) contribuyeron con normas específicas de edad basadas en 530 personas cognitivamente normales de 56 a 97 años de edad. Schmidt (1996) ha recopilado, resumido y sintetizado las normas disponibles para la RAVLT.

VALORACIÓN FULD DE MEMORIA DE OBJETOS

La *Fuld Object-Memory Evaluation* (Valoración Fuld de Memoria de Objetos) es una prueba útil del deterioro de la memoria en ancianos (Fuld, 1977). La prueba comienza con la presentación ante la persona evaluada de una bolsa que contiene 10 objetos comunes (pelota, botella, botón, etc.). La tarea no se describe como una prueba de memoria. Se pide al individuo evaluado que determine si puede identificar los objetos sólo con el tacto. Se siente cada uno y después se le nombra; posteriormente la persona extrae el objeto de la bolsa para ver si está en lo correcto. Después de que se han identificado de manera correcta los 10 artículos, se aplica una prueba de distracción: decir rápidamente palabras en una categoría semántica (p. ej., nombres, comidas, cosas que hacen felices a las personas, verduras o cosas que hacen entristecer a las personas). A continuación se pide a la persona evaluada que recuerde tantos objetos como sea posible. Después de cada objeto rememorado, se le recuerda de manera verbal al sujeto, en forma lenta y clara, cada uno de los artículos omitidos en ese ensayo, un proceso denominado recordatorio selectivo (Buschke y Fuld, 1974). Después se aplican a la persona cuatro oportunidades adicionales de rememorar la lista a través de recordatorio selectivo, con una tarea de distracción después de cada ensayo. El recuerdo demorado se evalúa después de un intervalo de cinco minutos. Por último, la prueba se cierra con una prueba de reconocimiento de opción múltiple.

Es frecuente que la prueba Fuld se utilice para ayudar a confirmar el diagnóstico de **enfermedad de Alzheimer**, trastorno neurológico degenerativo que se describió en el tema anterior. En las etapas iniciales de esta enfermedad, el síntoma más destacado es la pérdida de memoria. Las personas ancianas con deterioro en la memoria no sólo obtienen puntuaciones más bajas que los sujetos control en la Evaluación Fuld de memoria de objetos, sino que también se benefician muy poco del recordatorio selectivo. Fuld (1977) ha proporcionado normas para personas en los rangos de 70 a 90 años, activas en su comunidad y residentes sanos de asilos. Fuld, Masur, Blau, Crystal y Aronson (1990) describen un estudio prospectivo en el que la Evaluación Fuld de memoria de objetos demostró ser prometedora como medio de predicción de la demencia en ancianos cognitivamente normales. Lichtenberg, Manning, Vangel y Ross (1995) describen un programa de investigación neuropsicológica en el que se utiliza la prueba Fuld con pacientes médicos urbanos de edad avanzada.

PRUEBAS ADICIONALES DE APRENDIZAJE Y MEMORIA

Debido a las limitaciones de espacio, no puede hacerse más que una mención breve de otras pruebas útiles de aprendizaje y memoria. La *California Verbal Learning Test* (Prueba California de Aprendizaje Verbal) está diseñada con base en la AVLT de Rey, pero proporciona *software* de computadora para cuantificar y analizar el patrón de resultados (Schear y Craft, 1989). La *Benton Visual Retention Test* (Prueba Benton de Retención Visual) es una prueba de memoria visual con copia de diseños (Sivan, 1991). La *Rivermead Behavioral Memory Test* (Prueba Rivermead de memoria conductual) es una medida de la Memoria Cotidiana (p. ej., encontrar rutas, recordar un nombre) dentro de entornos de rehabilitación (Koltai, Bowler y Shore, 1996). En Lezak (1995), Reeves y Wedding (1994) y Spreen y Strauss (1988) pueden encontrarse buenas reseñas de pruebas de memoria.

EVALUACIÓN DE LAS FUNCIONES DEL LENGUAJE

Como se señaló en la sección anterior, el funcionamiento del lenguaje ofrece una ventana a la integridad del hemisferio cerebral izquierdo. Así, los neuropsicólogos están profundamente interesados en la capacidad de la persona eva-

luada para hablar, leer, escribir y comprender lo que otros dicen. Provoca poca sorpresa que una evaluación neuropsicológica completa siempre incluya uno o más métodos para evaluar las funciones del lenguaje.

Los neuropsicólogos exhiben un interés particular en una variedad de disfunciones del lenguaje conocidas de manera colectiva como afasia. En pocas palabras, la **afasia** es cualquier desviación en el desempeño del lenguaje producida por daño cerebral. En el examen de ésta, un neuropsicólogo puede utilizar cualquiera o todos los siguientes tres enfoques: 1) un examen clínico no estandarizado, 2) una prueba estandarizada para detección o 3) una prueba diagnóstica completa de la afasia. En nuestra breve revisión de los métodos de evaluación de la afasia se proporcionarán ejemplos de cada uno de estos enfoques.

EXAMEN CLÍNICO DE LA AFASIA

Un examen clínico de la afasia tiene las ventajas de la sencillez, flexibilidad y brevedad. Estos son atributos importantes cuando se examina a pacientes con deterioro grave que pueden requerir evaluación en su lecho de enfermos. Todo profesional tiene una versión diferente del examen clínico breve (Lezak, 1995; Reitan, 1984, 1985). No obstante, se evalúan ciertos elementos en común:

- **Habla espontánea**: el examinador busca síntomas distintivos de afasia como dificultad para encontrar palabras o neologismos (p. ej., referirse a un peine como "planquer").
- **Repetición de oraciones y frases**: el examinador le pide al paciente que repita estímulos como "por donde pecas, pagas" y "Metodista Episcopal". Las tareas de repetición son tan sencillas que las personas normales casi nunca las fallan.
- **Comprensión del lenguaje escrito**: el examinador hace preguntas ("¿Un automóvil tiene manijas?") y da órdenes ("Tome este papel, dóblelo a la mitad y colóquelo en el suelo"). De nuevo, las tareas son tan sencillas que los individuos normales nunca las fallan.
- **Localización de palabras:** el examinador señala a objetos comunes y fácilmente reconocibles y pregunta: "¿Qué es esto?" Reactivos típicos incluyen un reloj, pluma, lápiz, anteojos, anillo y zapatos. El examinador le puede pedir al paciente que nombre números, letras o colores.
- **Lectura:** el examinador le solicita al paciente que lea y explique un párrafo corto adecuado para el nivel previo de educación e inteligencia. Le puede pedir al paciente que siga instrucciones por escrito (p. ej., "Cierre los ojos" o "Aplauda tres veces con las manos").
- **Escritura y copiado**: el examinador le pide al paciente que escriba de manera espontánea o a partir de un dictado. También, le puede pedir al paciente que copie material escrito y formas geométricas. El examinador está interesado en producciones escritas con gramática sumamente errónea y distorsiones importantes en el copiado.
- **Cálculo**: el examinador le pide al paciente que lleve a cabo cálculos matemáticos sencillos (p. ej., 17×3) con y sin ayuda de papel para apuntes. Las tareas son tan simples que las personas normales pocas veces fallan.

Con base en la evaluación clínica, el examinador puede llenar una escala de calificación para la gravedad de la afasia. Por ejemplo, la escala de calificación que se utiliza en el *Boston Diagnostic Aphasia Exam* (Examen Boston de Diagnóstico de Afasia; Goodglass y Kaplan, 1983) incluye las siguientes características del habla: línea melódica, extensión de la frase, agilidad de la articulación, forma gramatical, localización de palabras y comprensión auditiva.

PRUEBAS DE DETECCIÓN Y DE DIAGNÓSTICO COMPLETO DE LA AFASIA

Las pruebas estandarizadas de detección de la afasia se asemejan de manera estrecha al examen clínico breve. La diferencia esencial es que

las pruebas estandarizadas de detección incorporan instrucciones objetivas y precisas para aplicación y calificación. La debilidad de las pruebas de detección es que no detectarán las formas sutiles de la afasia. En la figura 9–8 se presentan los estímulos para una prueba ampliamente utilizada para la detección de la afasia.

Las pruebas para el diagnóstico completo de la afasia son bastante extensas y se utilizan principalmente cuando se sabe que un paciente tiene afasia. Estas pruebas proporcionan un perfil de habilidades del lenguaje que es útil para la planificación del tratamiento. En el cuadro 9–10 se presenta una descripción breve de varias pruebas de afasia.

PRUEBAS DE CAPACIDAD ESPACIAL Y DE MANIPULACIÓN

Las pruebas de capacidad espacial y de manipulación también se conocen como pruebas de desempeño constructivo. Una prueba de este tipo combina la actividad perceptual con la respuesta motora y siempre tiene un componente espacial (Lezak, 1995). Debido a que la capacidad constructiva implica varias funciones complejas, incluso las formas leves de disfunción cerebral tendrán por resultado un deterioro en el desempeño constructivo. No obstante, se requiere de observación cuidadosa para distinguir las causas del desempeño erróneo, que pueden incluir confusión espacial, deficiencia perceptual, dificultades de atención, problemas de motivación y apraxias. El término **apraxia** se refiere a una variedad de disfunciones caracterizadas por una desintegración en la dirección o ejecución de actos motores complejos (Strub y Black, 1985). Por ejemplo, un paciente que no puede demostrar cómo se utiliza una llave recibiría un diagnóstico de apraxia ideomotora.

Las pruebas del desempeño constructivo engloban dos clases amplias de actividades: dibujo y ensamblado. Debido a las limitaciones de espacio, se revisarán sólo unos cuantos instrumentos destacados dentro de cada categoría.



Nota: Las tareas implican el establecimiento de nombres, deletreo, lectura, repetición y cálculo.

Figura 9–8. Figuras estímulo de la *Reitan-Indiana Aphasia Screening Test* (Prueba Reitan-Indiana de Detección de la Afasia).
Fuente: Reimpreso con autorización de Reitan, R. M. y Wolfson, D. (1993). *The Halstead-Reitan Neuropsychological Test Battery: Theory and Clinical Interpretation* (2ª Edición). Tucson, AZ: Neuropsychology Press. The Reitan-Indiana Aphasia Screening Test está disponible en Reitan Neuropsychology Laboratories, 2920 S. Fourth Ave., South Tucson, AZ 85713.

Cuadro 9–10. Descripción breve de varias pruebas de la afasia

***Multilingual Aphasia Examination* (Examen Multilingüe de la Afasia)**
Benton, Hamsher, Rey y Sivan (1994)
Esta batería respetada y amplia consiste de 11 subpruebas y escalas de calificación que evalúan nombramiento visual, repetición, fluidez, articulación, ortografía y otras variables del lenguaje; también disponible en español.

***Western Aphasia Battery* (Batería Occidental de Afasia)**
Kertesz (1979)
Prueba amplia de la fluidez verbal, comprensión auditiva y repetición que pretende identificar síndromes de afasia y determinar su gravedad.

***Boston Diagnostic Aphasia Examination* (Examen Boston de Diagnóstico de Afasia)**
Goodglass y Kaplan (1983)
Prueba amplia con 46 subescalas que incluyen habilidades musicales, espaciales, de cálculo y siete tipos de habilidades en escritura, además de medidas tradicionales de la afasia; disponible también en francés e hindi.

***Aphasia Screening Test* (Prueba de Detección de la Afasia)**
Reitan (1984, 1985)
Utiliza el enfoque de signo patognomónico (error = síntoma de afasia) y no produce un rango de puntuaciones. Es una buena prueba de detección pero es posible que no detecte deterioros sutiles.

***Aphasia Language Performance Scales*, ALPS (Escalas de Afasia de Desempeño en el Lenguaje)**
Keenan y Brassell (1975)
Se someten a prueba las cuatro principales modalidades del lenguaje (habla, audición, lectura, escritura) con 10 reactivos de complejidad creciente. Tiene buena confiabilidad (rangos de .80 y .90) y correlaciones robustas (.94) con el PICA, listado adelante.

***Porch Index of Communicative Ability*, PICA (Índice Porch de Habilidad de Comunicación),**
Porch (1967)
Una batería que contiene 18 subpruebas con 10 reactivos, cuatro verbales, ocho por ademanes y seis gráficas. Es una prueba muy confiable que se utiliza con frecuencia para medir pequeños cambios en el desempeño del paciente.

Cuadro 9–10. Descripción breve de varias pruebas de la afasia (continuación)

***Token Test* (Prueba de Fichas)**
Spreen y Strauss (1998)
Una prueba sumamente sensible que representa un desafío menor para los individuos normales. La persona evaluada debe llevar a cabo órdenes orales con fichas de colores; por ejemplo, "Coloque la ficha roja pequeña sobre la ficha cuadrada grande." Diseñada originalmente por Boller y Vignolo (1966), en la actualidad se encuentran disponibles numerosas versiones de la Prueba de Fichas.

PRUEBAS DE DIBUJO

Más allá de cualquier duda, la prueba de dibujo de más amplia utilización es el Test Guestáltico Visomotor, conocido de manera más común como Prueba Gestalt de Bender (BGT, Bender, 1938). La BGT consta de nueve figuras estímulo (figura 9–9); se instruye a la persona evaluada para que copie cada figura en una hoja de papel en blanco. Se dice al examinando que la BGT: "... no es una prueba de capacidad artística, pero intente copiar el dibujo lo más precisamente posible. Trabaje con la rapidez o lentitud que usted desee" (Hutt, 1977).

Se han diseñado varios sistemas de calificación para determinar si el desempeño de una persona es el más común en individuos con daño cerebral o sin él (Hain, 1964; Hutt y Briskin, 1960; Lacks, 1999; Pascal y Suttell, 1951; Pauker, 1976). En el caso de adultos, el mejor de estos enfoques de calificación se encuentra en Lacks (1999). Esta autora identificó 12 signos cualitativos que se califican según presencia contra ausencia en el protocolo completo. La presencia de cualesquiera de los cinco signos indica daño cerebral (cuadro 9–11). Con base en la confirmación independiente de otras fuentes de información, Lacks reseña tasas de éxito de 82 a 86% en una muestra combinada de admisiones a la unidad de tratamiento psiquiátrico agudo de un centro urbano de salud mental comunitaria (Lacks y Newport, 1980). En Gregory (1999) se analizan diversas varia-

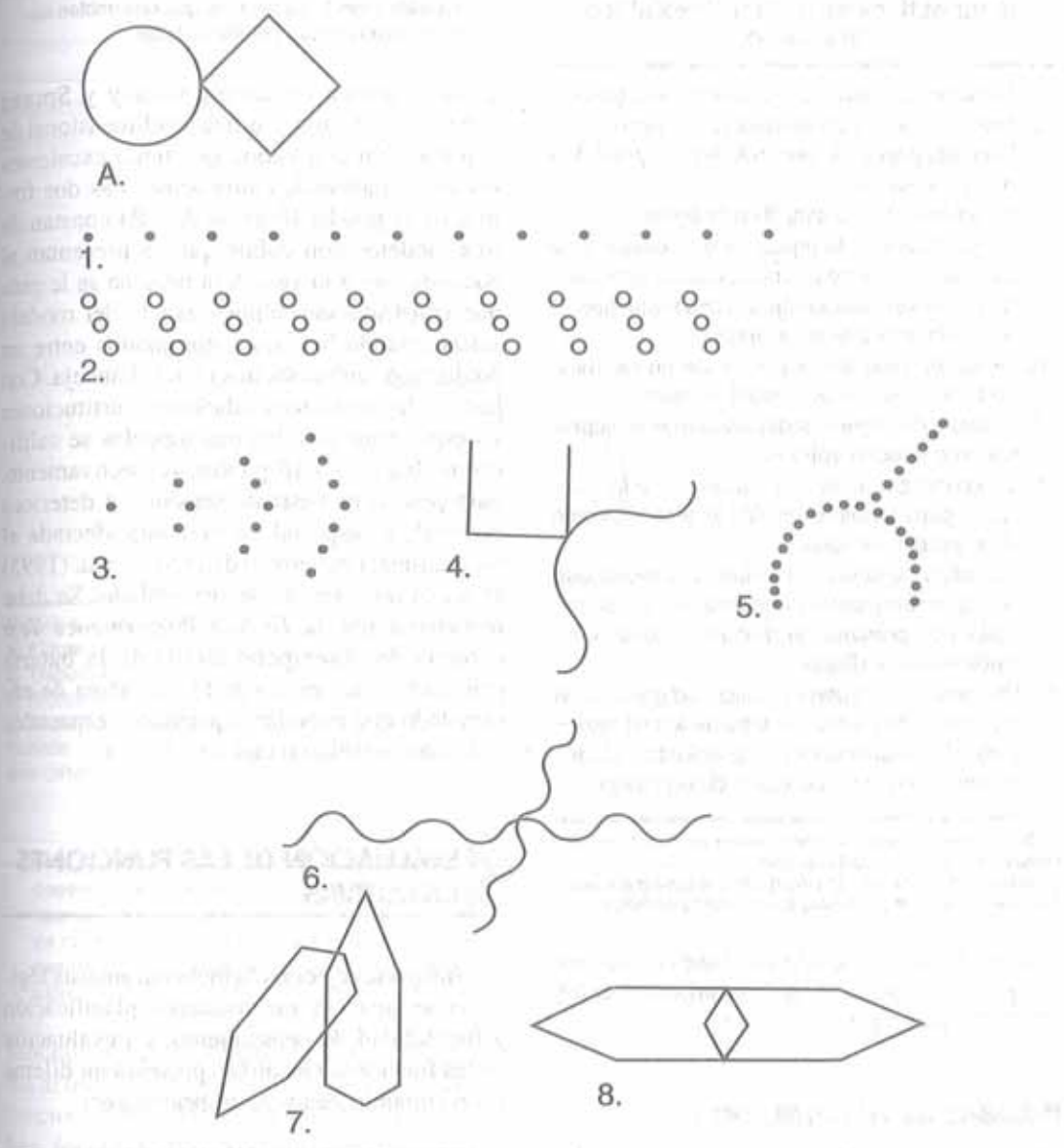


Figura 9–9. Estímulos para el Test Guestáltico Visomotor.
Fuente: Reproducido con autorización de Bender, L. (1938). *A visual motor gestalt test and its clinical use.* Nueva York: American Orthopsychiatric Association. Copyright © American Orthopsychiatric Association.

ciones interesantes en la BGT. Groth-Marnat (1990) dedica un capítulo completo a este instrumento, incluyendo pautas interpretativas para niños y adultos.

La Cruz Griega (Reitan y Wolfson, 1993) es una tarea muy sencilla de dibujo sorprendentemente sensible al deterioro cerebral. Se le solicita de manera cuidadosa a la persona evaluada que copie la figura sin levantar el lápiz, es decir, trazando el perímetro. En la figura 9–10 se muestran la figura estímulo y ejemplos de desempeño defectuoso. Es más frecuente que

Cuadro 9–11. Resumen de signos diagnósticos en el Test Guestáltico Visomotor

1. *Rotación*: la figura se rota de 80 a 180 grados.
2. *Dificultad de superposición:* problema para dibujar las partes de una sola figura que deberían traslaparse.
3. *Simplificación*: se simplifica la figura.
4. *Fragmentación*: la figura se descompone de tal manera que se pierde la gestalt general.
5. *Retrogresión*: sustitución con una forma *gestalt* más primitiva que el estímulo.
6. *Perseveración*: los aspectos de un estímulo anterior se llevan al estímulo actual.
7. *Colisión*: dos figuras independientes se superponen o chocan entre sí.
8. *Impotencia*: numerosas borraduras e incapacidad para terminar un dibujo a satisfacción de la propia persona.
9. *Dificultad de cierre*: dificultad para lograr que se toquen las partes adyacentes de una figura.
10. *Falta de coordinación motora:* el temblor es evidente en el dibujo.
11. *Dificultad de angulación:* dificultad grave en la reproducción de los ángulos de los dibujos.
12. *Cohesión:* disminución o incremento aislados en tamaño de una subparte de un dibujo.

Nota: Se puede contar un decimotercer error si la prueba completa requiere más de 15 minutos.
Fuente: Basado en Lacks, P. (1999). *Bender-Gestalt screening for brain dysfunction* (2ª edición). Nueva York: John Wiley.

esta prueba se valore con una base cualitativa, aunque existen pautas de calificación (Swiercinsky, 1978; Gregory, 1999).

PRUEBAS DE ENSAMBLADO

En su libro clásico sobre los lóbulos parietales, Critchley (1953) proporcionó la fundamentación para incluir tareas de construcción tridimensional en una batería de pruebas neuropsicológicas:

> Es posible, y de hecho es útil, proceder a problemas en espacio tridimensional, aunque las pruebas de este tipo sólo se emplean en raras ocasiones. Ésta es una tarea más difícil y los pacientes que responden moderadamente bien a los procedimientos comunes con varitas y lápiz y papel, pueden presentar anormalidades notables cuando se les pide que ensamblen cubos según un patrón tridimensional.

Benton, Sivan, Hamsher, Varney y Spreen (1994) presentan una prueba tridimensional de construcción con cubos que tiene excelentes normas y pautas de calificación. Las dos formas de la prueba (Formas A y B) constan de tres modelos con cubos que se presentan al paciente uno a la vez. A la persona se le pide que construya una réplica exacta del modelo seleccionando los cubos apropiados entre un conjunto de cubos sueltos en una bandeja. Con base en las omisiones, adiciones, sustituciones y desplazamientos, los tres modelos se califican de 0 a 6, 8 y 15 puntos, respectivamente. Esta prueba es bastante sensible al deterioro cerebral, en especial cuando está afectada el área parietal izquierda o derecha. Lezak (1995) examina otras tareas de ensamblado. Se debe mencionar que la *Tactual Performance Test* (Prueba de desempeño táctil) de la batería Halstead-Reitan es, en parte, una tarea de ensamblado que mide las capacidades espaciales y de manipulación (cuadro 9–7).

EVALUACIÓN DE LAS FUNCIONES EJECUTIVAS

Las **funciones ejecutivas** incluyen análisis lógico, conceptuación, razonamiento, planificación y flexibilidad de pensamiento. La evaluación de las funciones ejecutivas presenta un dilema poco común para los neuropsicólogos:

> Uno de los principales obstáculos para examinar las funciones ejecutivas es la paradójica necesidad de estructurar una situación en la que los pacientes puedan mostrar si pueden y qué tan bien pueden conformar una estructura por sí mismos. Es común que en los exámenes formales el examinador determine cuál actividad debe realizar el individuo y con qué materiales, cuándo, dónde y cómo. Por ejemplo, la mayor parte de las pruebas cognitivas le dan poca oportunidad a la persona de presentar conducta discrecional, incluyendo muchas pruebas que se consideran sensibles a los trastornos ejecutivos —o del ló-

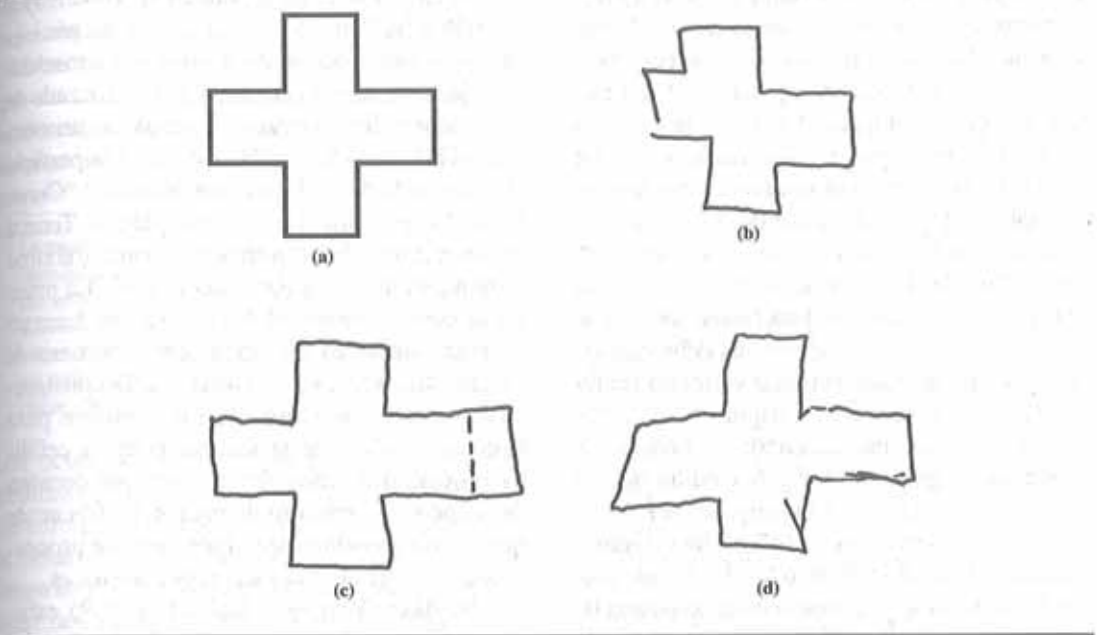


Figura 9–10. Figura estímulo de la Cruz Griega y reproducciones de personas con daño cerebral conocido.
a) Figura estímulo.
b) Trabajador secretarial con disfunción difusa del hemisferio derecho de origen desconocido.
c) Profesor universitario dos años después de una embolia en el hemisferio derecho.
d) Paciente con demencia difusa, generalizada.
Fuente: Reproducido con autorización de Gregory, R. J. (1999). *Foundations of intellectual assessment: The WAIS-III and other tests in clinical practice.* Boston: Allyn & Bacon.

> bulo frontal... El problema para los clínicos que desean examinar las funciones ejecutivas tiene que ver con el cómo transferir el establecimiento de metas, la estructuración y la toma de decisiones del clínico al individuo dentro de un examen estructurado (Lezak, 1995).

Muchos neuropsicólogos resuelven este dilema al utilizar el método clínico para evaluar las funciones ejecutivas en lugar de aplicar pruebas formales (Cripe, 1996). Por ejemplo, Pollens, McBratnie y Burton (1988) utilizan una entrevista y observaciones para llenar la lista estructurada de verificación sobre funciones ejecutivas que se menciona en el tema anterior.

Sólo un número limitado de pruebas neuropsicológicas detectan las funciones ejecutivas en grado razonable. Instrumentos útiles en este sentido incluyen el Test de Laberintos Porteus, la *Wisconsin Card Sorting Test* (Prueba Wisconsin de Clasificación de Tarjetas) y un enfoque novedoso conocido como *Tinkertoy® Test* (Prueba Tinkertoy®). Se le recuerda al lector que la Prueba de categorías de la batería Halstead-Reitan también capta las funciones ejecutivas hasta cierto nivel (cuadro 9–7).

El Test de Laberintos Porteus se diseñó como una medida con reducción cultural de la planificación y previsión (Porteus, 1965). Sin levantar el lápiz e intentando evitar los pasadizos sin salida, la persona evaluada debe trazar una línea a través de una serie de laberintos de dificultad cada vez mayor. Este instrumento que se utiliza muy poco es bastante sensible a los efectos del daño cerebral, en particular en los lóbulos frontales (Smith y Kinder, 1959; Smith, 1960; Tow, 1955).

La *Wisconsin Card Sorting Test* (Prueba Wisconsin de clasificación de tarjetas) es una buena medida de las funciones ejecutivas, aunque se debate su sensibilidad diferencial al daño

en el lóbulo frontal (Mountain y Snow, 1993). El instrumento se diseñó para estudiar el pensamiento abstracto y la capacidad para cambiar de propensión (Berg, 1948; Heaton, Chelune, Talley y colaboradores, 1993). A la persona evaluada se le proporciona un paquete con 64 tarjetas en las que están impresos uno de cuatro símbolos (triángulo, estrella, cruz y círculo) en uno de cuatro colores (rojo, verde, amarillo o azul). No hay dos tarjetas idénticas. Así, cada una incluye un número, una forma particular y un color específico. La persona debe clasificarlas debajo de cuatro tarjetas estímulo según un principio desconocido (figura 9–11). Por ejemplo, el principio desconocido podría ser "clasifique según el color". A medida que la persona evaluada coloca las tarjetas, el examinador puede decir "correcto" o "incorrecto". Después de que el individuo ha clasificado una secuencia de 10 asignaciones correctas en una fila, el examinador cambia el principio sin advertencia. La prueba continúa hasta que la persona ha hecho seis secuencias de 10 asignaciones correctas. La prueba puede calificarse de diferentes maneras, incluyendo el número total de ensayos para llegar al criterio (Axelrod, Greve y Goldman, 1994).

Lezak (1982) ha diseñado la *Tinkertoy® Test* (Prueba Tinkertoy®) para dar a los pacientes la oportunidad de demostrar sus capacidades ejecutivas dentro del formato estructurado de un examen. Se colocan 50 piezas de un conjunto *Tinkertoy®* estándar sobre una superficie limpia y se le dice a la persona evaluada: "Construya lo que quiera con estas piezas. Tendrá como mínimo cinco minutos y como máximo el tiempo que desee para hacer algo". La prueba se califica desde –1 hasta +12 con base en diversas variables que incluyen el número de piezas utilizadas, la movilidad de la construcción, simetría y asignación de nombre para la construcción. Los pacientes con lesión cefálica producen diseños deficientes que constan de un pequeño número de piezas. Es frecuente que estos individuos sean incapaces de proporcionar un nombre para sus construcciones.

Bayless, Varney y Roberts (1989) estudiaron la validez predictiva de la Prueba Tinkertoy® al comparar los resultados de 50 pacientes con lesiones cefálicas cerradas contra 25 controles normales. La mitad de los individuos con este tipo de lesiones había regresado a su trabajo y la mitad no lo había hecho. Mientras que todos, menos uno, de los pacientes

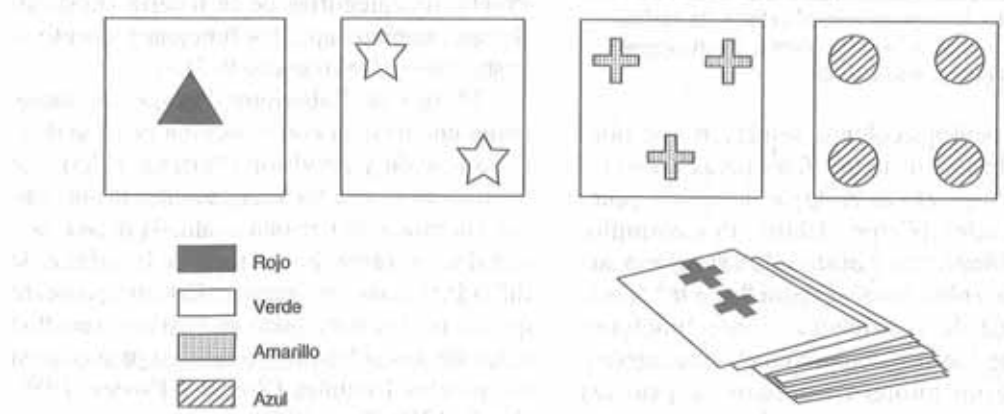


Figura 9–11. Tarjetas y pilas de clasificación para la *Wisconsin Card Sorting Test* (Prueba Wisconsin de clasificación de tarjetas).
Fuente: Reproducido con autorización especial de Psychological Assessment Resources, Inc. Derechos reservados. Copyright © 1981 de Psychological Assessment Resources, Inc.

con lesión cefálica que habían regresado al trabajo obtuvieron puntuaciones normales en la prueba Tinkertoy®, casi la mitad de aquellos que no habían regresado a su trabajo tuvieron un desempeño por debajo del nivel del peor participante control. Los investigadores concluyeron que:

> La prueba parece particularmente adecuada para demostrar la presencia de deficiencias en el funcionamiento ejecutivo, las cuales han resultado difíciles de demostrar con pruebas clínicas, a pesar de que tienen secuelas catastróficas en los esfuerzos cotidianos vocacionales o psicosociales (Bayless y cols., 1989).

Los neuropsicólogos continúan necesitando medidas adicionales de las funciones ejecutivas. Un enfoque prometedor en las primeras etapas de desarrollo es la evaluación realista del hallazgo de rutas. La capacidad para encontrar un lugar desconocido en una ciudad requiere de estrategia, automonitoreo y maniobras correctivas. Éstas son funciones ejecutivas que se aplican a un problema real (Boyd y Sauter, 1993). Otro enfoque prometedor se encuentra en una reciente batería denominada *Behavioral Assessment of Dysexecutive System* (BADS) (Evaluación conductual del sistema disejecutivo; Wilson, Alderman, Burgess y cols., 1996). La batería BADS consta de seis nuevas pruebas que son similares a las actividades de la vida real:

1. Cambio de reglas en juegos de naipes
2. Solución de problemas con un vaso de precipitados lleno de agua
3. Búsqueda simulada de una llave perdida
4. Juicio del tiempo necesario para actividades, acontecimientos
5. Hallazgo de ruta con un mapa
6. Tarea de organización que implica dictado, aritmética, asignación de nombre

La batería también incluye un cuestionario de 20 reactivos que se califica en una escala Likert de cinco puntos (0 a 4). Los reactivos incluyen cambios probables cuando existe deterioro en las funciones ejecutivas, por ejemplo, "Tengo dificultad para pensar de antemano y planear para el futuro". Spreen y Strauss (1998) proporcionan una útil reseña sobre esta prometedora batería de pruebas.

EVALUACIÓN DE LA SALIDA MOTORA

La mayor parte de las baterías de pruebas neuropsicológicas incluyen medidas de velocidad y precisión en la manipulación. Lezak (1995) proporciona una reseña amplia. Aquí se resumirán brevemente tres enfoques: golpeteo con los dedos, desempeño en un tablero de clavijas y trazado de líneas.

Quizá la prueba de uso más amplio de la destreza motora es la *Finger-Tapping Test* (Prueba de Golpeteo con los Dedos) de la batería Halstead-Reitan. Ésta consta de una tecla para oprimir que sale de un instrumento mecánico de conteo incorporado dentro de un tablero plano. Con el dedo índice de cada mano, la persona evaluada lleva a cabo una serie de ensayos de 10 segundos hasta que cinco seguidos se encuentran dentro de un rango de cinco puntos. La puntuación para cada mano es el promedio de estos cinco ensayos, redondeada al número entero más cercano. Con la mano dominante, los varones por lo común alcanzan puntuaciones de 54 golpes (DE de 4), mientras las mujeres suelen tener puntuaciones de cerca de 51 golpes (DE de 5, Dodrill, 1979; Morrison, Gregory y Paul, 1979).

En general, el nivel absoluto de desempeño es de menor interés que las capacidades relativas en ambos lados del cuerpo. La expectativa normativa es que la mano no dominante tendrá una tasa de golpeteo de aproximadamente 90% con relación a la dominante. Se considera que las desviaciones significativas con respecto a este patrón indican una lesión en el hemisferio opuesto a la mano más lenta (Haaland y Delaney, 1981). No obstante, dichas inferencias deben realizarse con gran cuidado, debido a la baja confiabilidad de la puntuación de proporción. Aunque las confiabilidades *test-retest* e

interexaminadores para cada mano se acercan a .80, la confiabilidad de la puntuación de proporción es un pésimo .44 a .54 (Morrison, Gregory y Paul, 1979). La puntuación de proporción debe utilizarse con extremo cuidado cuando se hacen inferencias clínicas sobre la lateralización del daño.

El Test no verbal de la Universidad de Purdue requiere que la persona evaluada coloque clavijas dentro de orificios con la mano izquierda, derecha y después con ambas manos. Cada ensayo dura sólo 30 segundos, de modo que la prueba completa puede aplicarse en cuestión de minutos. Tiffin (1968) informa de puntuaciones normativas para solicitantes de empleo. La lentitud relativa en una mano sugiere una lesión en el hemisferio opuesto, mientras la lentitud bilateral indica daño cerebral difuso o bilateral. Al utilizar el Test no verbal de la Universidad de Purdue de manera aislada, un estudio encontró 80% de precisión en la identificación de deterioro cerebral en un gran grupo de individuos normales y pacientes neurológicos (Lezak, 1983). Otros estudios informan datos menos favorables (Heaton, Smith, Lehman y Vogt, 1978). El Test no verbal de la Universidad de Purdue es una adición útil a la batería completa, pero no debe utilizarse por sí solo para propósitos de detección. Spreen y Strauss (1998) proporcionan un excelente resumen de las normas de esta prueba ampliamente utilizada.

Klove ha desarrollado una variación de la prueba de tablero de clavijas en la que éstas tienen un borde en uno de los lados (Klove, 1963). Debido a que cada clavija tiene que girarse a una posición, el tablero perforado para clavijas requiere de coordinación compleja, además de destreza motora. La *Grooved Pegboard Test* (Tablero Perforado para Clavijas) es un excelente instrumento para evaluar el daño cerebral lateralizado (Haaland y Delaney, 1981).

Por último, debe mencionarse que las pruebas motoras útiles no requieren de equipo complejo. Lezak (1995) recomienda una *Line Tracing Task* (Tarea de Trazado de Líneas) para evaluar las dificultades en la regulación motora (figura 9–12). Se proporciona al examinando una pluma de colores brillantes con punta de fieltro y una hoja de papel con varias figuras y se le pide que dibuje sobre las líneas con la mayor rapidez posible. Las dificultades con la regulación motora se observan en el desbordamiento en las esquinas, perseveración en una respuesta continua e incapacidad para seguir las curvas reducidas en la última figura. Debido a que esta tarea la llevan a cabo con facilidad la mayoría de los niños de 10 años, cualesquiera desviaciones notables sugieren dificultades en regulación motora.

BATERÍAS DE PRUEBA EN EVALUACIÓN NEUROPSICOLÓGICA

Ahora que se ha completado un recorrido sobre algunas pruebas y procedimientos neuropsicológicos individuales, es tiempo otra vez de recordar al lector que muchos neuropsicólogos prefieren utilizar una batería fija en vez de un surtido siempre cambiante e individualizado de instrumentos. Con toda seguridad una de las baterías fijas de más amplia utilización es el *Luria-Nebraska Neuropsychological Battery*, LNNB (Diagnóstico Neuropsicológico Luria-Nebraska; Golden, 1989; Golden, Purish y Hammeke, 1980, 1986).

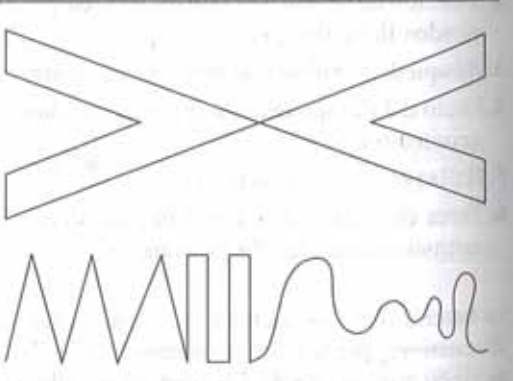


Figura 9–12. Tarea típica de trazado de líneas (tamaño reducido).

La prueba consta de 269 reactivos separados, elegidos del trabajo de Luria (p. ej., 1966) y estandarizados de manera formal. Estos reactivos se califican con 0, 1 o 2, según los criterios precisos del manual de aplicación y calificación. Los reactivos similares se encuentran agrupados en 11 escalas clínicas, de la C1 a la C11 (cuadro 9–12). Las puntuaciones naturales en cada escala se convierten a puntuaciones T, con media de 50 y desviación estándar de 10. Las puntuaciones más altas reflejan mayor psicopatología; aquellas superiores a 70 son especialmente sugerentes de daño cerebral.

Cuadro 9–12. Pruebas y procedimientos de la *Luria-Nebraska Neuropsychological Battery*

Escala de capacidad: Tareas incluidas
C1 Motora: coordinación, velocidad, dibujo, capacidades motoras complejas.
C2 Rítmica: atender a, discriminar y producir estímulos rítmicos verbales y no verbales.
C3 Táctil: identificar estímulos táctiles, incluyendo estímulos trazados sobre las muñecas.
C4 Visual: identificar dibujos, incluyendo objetos superpuestos o fuera de foco; resolver matrices progresivas y otras habilidades visoespaciales.
C5 Discurso receptivo: discriminar fonemas y comprender palabras, frases, oraciones.
C6 Discurso expresivo: articular sonidos, palabras y oraciones de manera fluida; identificar objetos representados en imágenes o descritos.
C7 Escritura: utilizar capacidades motoras de escritura en general; copiar y escribir a partir de dictado.
C8 Lectura: leer letras, palabras y oraciones; sintetizar las letras en sonidos y palabras.
C9 Aritmética: llevar a cabo cálculos matemáticos sencillos; comprender los signos matemáticos y estructura de los números.
C10 Memoria: recordar estímulos verbales y no verbales bajo condiciones tanto de interferencia como sin ella.
C11 Inteligencia: razonamiento, formación de conceptos y solución de problemas matemáticos complejos.

También se obtienen tres escalas de resumen a partir del desempeño en prueba; S1 (Patognomónica), S2 (Hemisferio izquierdo) y S3 (Hemisferio derecho). La primera escala refleja el grado de compensación que ha ocurrido desde la lesión, como la reorganización funcional del cerebro, así como la recuperación física en sí. Las puntuaciones más altas reflejan menor compensación. Las escalas de Hemisferio izquierdo y Hemisferio derecho pueden utilizarse para ayudar a determinar si una lesión es difusa o lateralizada. También están disponibles varias otras escalas y factores interpretativos (Golden, Purish y Hammeke, 1986). El empleo del LNNB se ilustra en el ejemplo de caso 9–2.

En este libro no es posible revisar la voluminosa literatura acerca del LNNB, no obstante con seguridad es necesario mencionar de modo breve unos cuantos estudios clave. La confiabilidad del LNNB se ha valorado a partir de las perspectivas comunes (por mitades, consistencia interna y *test-retest*), con resultados excelentes. Por ejemplo, la confiabilidad media *test-retest* para las escalas clínicas es de cerca de .90 (Bach, Harowski, Kirby, Peterson y Schulein, 1981; Plaisted y Golden, 1982). En diversos estudios de validez acerca de la clasificación de personas con daño cerebral contra otros grupos de criterio, el LNNB ha mostrado tasas de éxito de 80% o mejores (Golden, Moses, Graber y Berg, 1981; Hammeke, Golden y Purish, 1978; Moses y Golden, 1979).

A pesar de las conclusiones positivas que informaron Golden y sus colaboradores, muchos neuropsicólogos continúan mostrando escepticismo acerca del LNNB (Lezak, 1995; Kolb y Whishaw, 1990; Crosson y Warren, 1982; Delis y Kaplan, 1982, 1983). Por ejemplo, Delis y Kaplan (1983) afirmaron que la heterogeneidad de las escalas es tan grande que las puntuaciones no cuantifican deficiencias neuropsicológicas específicas:

> En la "validación" de la batería, la elevación en una escala se analizó sólo en términos de la diferenciación de un grupo de pacientes "con daño cerebral" con respecto a personas "normales". Nunca se tomó siquiera en consideración la

EJEMPLO DE CASO 9–2
DIAGNÓSTICO NEUROPSICOLÓGICO LURIA-NEBRASKA (LNNB)

Para ilustrar la aplicación del LNNB, se presenta el caso de un varón de 56 años de edad, con educación de posgrado (Golden, 1989). Este paciente experimentó dos embolias en el hemisferio derecho antes de la aplicación del LNNB. Sus puntuaciones en las escalas clínicas y de resumen se muestran abajo. El lector observará la marcada elevación de S3 (Hemisferio derecho, $T = 124$), en comparación con S2 (Hemisferio izquierdo, $T = 81$), lo cual señala una lesión que está fuertemente lateralizada hacia el lado derecho del cerebro. También véase que S1 presenta una elevación solamente moderada ($T = 75$), lo cual indica que el paciente ha logrado un progreso adecuado para compensar su deterioro cerebral. El lenguaje (C5 y C6) e inteligencia (C11) están intactos en términos generales, pero lectura, escritura y memoria (C7, C8 y C10) muestran un posible deterioro sutil (T = 60, 60 y 65, respectivamente). Las discapacidades más graves se presentan en C1 a C4, es decir, Motora, Rítmica, Táctil y Visual. Esto indica grave deterioro sensoriomotor. El examen de los errores específicos en las escalas reveló que la visión participó en 74% de los errores. El paciente mostró un descuido unilateral izquierdo consistente, en el que se ignoró el lado izquierdo de los estímulos visuales (como en palabras u oraciones). En general, los resultados son bastante consistentes con un accidente cerebral vascular mayor en la arteria cerebral media del hemisferio derecho. La embolia dañó aparentemente las áreas sensoriomotoras y también los haces neuronales subcorticales que terminan en las áreas visuales del cerebro.

Escala		Puntuación T
C1	Motora	99
C2	Rítmica	92
C3	Táctil	92
C4	Visual	85
C5	Receptiva	39
C6	Expresiva	40
C7	Escritura	60
C8	Lectura	60
C9	Aritmética	45
C10	Memoria	65
C11	Inteligencia	53
S1	Patognomónica	75
S2	Hemisferio izquierdo	81
S3	Hemisferio derecho	124

Nota: Las puntuaciones esperadas para personas normales son aproximadamente de 50; puntuaciones superiores reflejan mayor deterioro.

validez de constructo acerca de si la elevación en una escala particular representaba un daño en esa capacidad.

Los mismos autores demostraron que las escalas relacionadas con el habla no se orientaron a síndromes de afasia y, por tanto, es posible que dieran un diagnóstico incorrecto de las deficiencias en lenguaje (Delis y Kaplan, 1982). Lezak (1995) señala la brecha que separa a las valoraciones del LNNB realizadas por Golden y colaboradores de aquéllas hechas por otros neuropsicólogos. Con base en su lectura de la literatura, Kolb y Whishaw (1990) concluyen que la "utilidad y validez de esta batería está lejos de haberse probado". Snow (1992) y Van Gorp (1992) también recomiendan precaución, al señalar que el LNNB es adecuado para identificar a individuos con deterioro cerebral en general, pero es cuestionable en la identificación del tipo, localización y consecuencias del daño cerebral. Otro problema con el LNNB es la alta tasa de falsos positivos en el caso de pacientes geriátricos. Con personas normales de 70 años de edad y mayores, las tasas de falsos positivos van de 21 a 41%, dependiendo del método de interpretación que se utilice (Roecker, House y Graybill, 1992-93).

EVALUACIÓN DEL ESTADO MENTAL EN ANCIANOS

El examen de *status mental* (ESM) es una entrevista aproximadamente estructurada que por lo general precede a otras formas de evaluación psicológica y médica. El propósito de ésta consiste en proporcionar una descripción precisa del funcionamiento del paciente en las áreas de orientación, memoria, pensamiento, sentimiento y juicio. El ESM es el equivalente psicológico del examen físico general: de la misma manera en que el médico revisa todos los principales sistemas orgánicos en busca de evidencia de enfermedad, el psicólogo revisa las principales categorías de funcionamiento personal e intelectual en búsqueda de signos y síntomas de psicopatología (Gregory, 1999). Aunque existe cierta libertad en cuanto al espectro del ESM, siempre se investigan ciertas funciones mentales. Una valoración típica toca las áreas listadas en el cuadro 9–13.

Cuadro 9–13. Principales áreas de un examen típico del estado mental

Apariencia y conducta
Arreglo personal
Expresiones faciales
Conducta motora gruesa
Contacto visual
Discurso y procesos de comunicación
Contenido, frecuencia, tono y volumen del discurso
Dificultad, confusión o empleo inadecuado de las palabras
Contenido del pensamiento
Lógica, claridad, idoneidad
Delirios
Funcionamiento cognitivo y de la memoria
Capacidad de cálculo
Recuerdo inmediato
Memoria reciente y remota
Reserva de información
Capacidad de abstracción
Funcionamiento emocional
Estado de ánimo predominante
Idoneidad del afecto
***Insight* y juicio**
Consciencia de los problemas
Orientación
Día, fecha, hora, lugar.

Fuente: Basado en Gregory, R. J. (1999). *Foundations of intellectual assessment: The WAIS-III and other test in clinical practice.* Boston, MA: Allyn y Bacon.

Algunos de los elementos de esta lista pueden evaluarse con pruebas cortas de detección. En particular, la cognición, memoria y orientación son funciones intelectuales que pueden examinarse de manera formal, estructurada (Hodges, 1994). En esta sección se revisarán varias medidas breves del estado mental que emplean los clínicos para complementar sus impresiones de entrevista. Estas medidas se emplean más en la valoración del estado mental de los ancianos, en especial cuando el cliente parece tener un tipo de demencia como la enfermedad de Alzheimer. Las pruebas formales del estado mental son útiles también en la evaluación de ciertos padecimientos con deterioro cerebral como lesiones cefálicas, esquizofrenia, depresión grave y delirio inducido por drogas. Es importante enfatizar que las pruebas de detección son complementarias —no reemplazan al juicio clínico en la valoración del estado mental. Algunas de las áreas que se cubren con el ESM son simplemente imposibles de cuantificar. Por ejemplo, la valoración del *insight* de un paciente requiere de observación aguda y de habilidades sensibles para la entrevista. No existe una prueba de detección para el ESM que mida el *insight*.

EXAMEN DEL ESTADO MINIMENTAL

La herramienta de uso más amplio para el estado mental es el *Mini-mental state examination*, MMSE (Examen del Estado Minimental), una prueba de detección de 5 a 10 minutos que produce un índice global del funcionamiento cognitivo (Folstein, Folstein y McHugh, 1975; Tombaugh, McDowell, Kristjansson y Hubley, 1996). La prueba contiene 30 reactivos calificables que tienen que ver con orientación, memoria inmediata, atención, cálculo, producción y comprensión del lenguaje y copia de diseños. Los reactivos son tan fáciles que la mayoría de los adultos normales casi siempre obtienen puntuaciones en el rango de 27 a 30 puntos (figura 9–13).

La confiabilidad de este sencillo instrumento es excelente. Folstein y colaboradores (1975) informan de una confiabilidad *test-retest* a 24 horas de .89 para 22 pacientes con diversos síntomas de depresión. La confiabilidad para un periodo de 28 días con 23 pacientes clínicamente estables con diagnósticos de demencia, depresión y esquizofrenia fue un impresionante .99. Se encuentran disponibles datos normativos que provienen de varias fuentes (p. ej., Lindal y Stefansson, 1993; Tombaugh, McDowell, Kristjansson y Hubley, 1996).

Con una puntuación de corte de 23 o menor como anormal y 24 o mayor como normal, el MMSE tiene cerca de 80 a 90% de precisión en la identificación de pacientes ancianos de los que se sospecha la presencia de enfermedad de Alzheimer u otros tipos de demencia. Esta puntuación de corte produce menos falsos positivos (pacientes normales a los que se clasifica dentro del diagnóstico de demencia). La sensibilidad del instrumento depende de varios factores, incluyendo la puntuación de corte utilizada, el nivel educativo de la persona evaluada, el grado de la demencia, la naturaleza de la patología subyacente y el tipo de entorno en el que se lleva a cabo la evaluación (Anthony, LeResche, Niaz, Von Korff y Folstein, 1982; Tombaugh, McDowell, Kristjansson y Hubley, 1996; Tsai y Tsuang, 1979). A pesar de sus limitaciones, el MMSE continúa siendo la prueba más confiable y práctica para la detección de demencia en ancianos (Ferris, 1992). Debring, Van Gorp, Stuck y otros investigadores (1994) recomiendan su empleo como parte de una batería corta de detección de la declinación cognitiva en ancianos. En el cuadro 9–14 se resumen medias adicionales del estado mental geriátrico.

MiniMental LLC

NOMBRE DEL PACIENTE _____ Edad _____

NOMBRE DEL EXAMINADOR _____ Años de estudio terminados _____

Acérquese al paciente con respeto y aliento — Fecha de evaluación _____

Pregunte: "¿Tiene algún problema con su memoria?" ☐ Sí ☐ No

"¿Puedo hacerle algunas preguntas sobre su memoria?" ☐ Sí ☐ No

CALIFICACIÓN	REACTIVO
5 ()	**ORIENTACIÓN EN TIEMPO** Pregunte: "¿Cuál es el año _____ (1), estación _____ (1), mes del año _____ (1), fecha _____ (1), día de la semana?" _____ (1)?"
5 ()	**ORIENTACIÓN EN LUGAR** Pregunte: "¿Dónde nos encontramos? ¿Cuál es el estado _____ (1), ciudad _____ (1), parte de la ciudad _____ (1), edificio _____ (1), piso del edificio?" _____ (1)?"
3 ()	**REGISTRO DE TRES PALABRAS** Diga: "Escuche cuidadosamente. Voy a decir tres palabras. Usted repítalas después de que me detenga. ¿Listo? Son las siguientes. . . POTRO (espere un segundo), CUARTO (espere un segundo), NARANJA (espere un segundo). ¿Cuáles eran esas palabras? _____ (1) _____ (1) _____ (1) Conceda un punto por cada respuesta correcta, después repítalas hasta que el paciente se aprenda las tres.
5 ()	**SIETES EN SERIE COMO PRUEBA DE ATENCIÓN Y CÁLCULO** Pregunte: "Reste 7 de 100 y continúe restando sietes de cada resultado subsecuente hasta que le pida que se detenga. ¿Cuánto es 100 menos 7?" _____ (1) Diga: "Prosiga". _____ (1), _____ (1), _____ (1), _____ (1).
3 ()	**RECUERDO DE TRES PALABRAS** Pregunte: "¿Cuáles fueron las tres palabras que le pedí que recordara?" Conceda un punto por cada respuesta correcta. _____ (1), _____ (1), _____ (1).

Para mayor información o copias adicionales de este examen, llame al (617) 587-4215
© 1998 M. FOLSTEIN

Figura 9–13. Examen del estado minimental.
Fuente: Folstein, M., Folstein, S. y McHugh, P. (1975). Mini-Mental State: A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 12*, 189-198. El copyright del Mini-Mental State Examination es propiedad completa de Mini Mental LLC, una compañía de responsabilidad limitada en Massachusetts. Para información acerca de la obtención de permisos para utilización o reproducción del Mini-Mental State Examination, por favor comunicarse con John Gonsalves, Jr., Administrador del Mini-Mental LLC, en el 31 St. James Avenue, Suite 1, Boston, MA 02116. (617) 587-4215. © 1998 MMLLC.

2 () **NOMBRES**

Pregunte:

"¿Qué es esto?" (muestre un lápiz) _____ (1). "¿Qué es esto?" (muestre un reloj) _____ (1).

1 () **REPETICIÓN**

Diga:

"Ahora le voy a pedir que repita lo que digo. ¿Listo? Por donde pecas, pagas."

"Ahora usted lo repite". _____ (1)

3 () **COMPRENSIÓN**

Diga:

"Escuche con cuidado porque voy a pedirle que haga una cosa:

"Tome este papel con su mano izquierda (1), dóblelo a la mitad (1) y colóquelo en el piso." (1)

1 () **LECTURA**

Diga:

"Por favor lea lo siguiente y haga lo que dice, pero no lo diga en voz alta."

Cierre los ojos

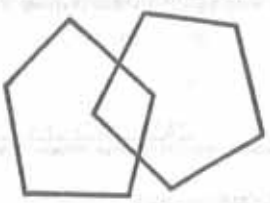
1 () **ESCRITURA**

Diga:

"Por favor escriba una oración." Si el paciente no responde, diga: "Escriba sobre el clima." (1)

1 () **DIBUJO**

Diga: "Por favor copie este diseño."



PUNTUACIÓN TOTAL _____ Evalúe el nivel de conciencia a lo largo de un continuo

Alerta — Somnolencia — Estupor — Coma

	SÍ	NO
Cooperativo:	☐	☐
Deprimido:	☐	☐
Ansioso:	☐	☐
Visión deficiente:	☐	☐
Audición deficiente:	☐	☐
Lengua materna:		

	SÍ	NO
Deterioro con respecto al nivel previo: de funcionamiento:	☐	☐
Antecedentes familiares de demencia:	☐	☐
Trauma encefálico:	☐	☐
Accidente cerebral vascular:	☐	☐
Abuso de alcohol:	☐	☐
Enfermedad de la tiroides:	☐	☐

FUNCIÓN POR PODER

Por favor registre los datos cuando el paciente fue menos capaz de llevar a cabo las siguientes tareas. Pregunte a la persona encargada del cuidado si el paciente maneja de manera independiente:

	SÍ	NO	FECHA
Dinero/billetes:	☐	☐	____
Medicamento:	☐	☐	____
Transporte:	☐	☐	____
Teléfono:	☐	☐	____

Figura 9–13. (continuación).

Cuadro 9–14. Pruebas de estado mental utilizadas con pacientes geriátricos

Prueba	Contenido
Cognistat Kiernan, Mueller y Langston (1997)	Lenguaje, construcción (copias), memoria, cálculo y razonamiento/juicio
Information-Memory-Concentration (Información-Memoria-Concentración) Blessed, Tomlinson y Roth (1968)	Información, orientación, concentración
Short Portable Mental Status Questionnaire (Cuestionario Portatil Breve del Estado Mental) Pfeiffer (1975)	Información, orientación, atención
Dementia Rating Scale (Escala de calificación de la demencia) Mattis (1976)	Atención, memoria, construcción (copias), conceptuación, fluidez verbal
Test of Temporal Orientations (Prueba de orientaciones temporales) Benton, Sivan, Hamsher, Varney y Spreen (1994)	Orientación
Alzheimer's Disease Assessment Scale (Escala de Evaluación de la Enfermedad de Alzheimer) Rosen, Mohs y Davis (1984)	Orientación, memoria, lenguaje, construcción (copias)
Cambridge Cognitive Examination (Examen Cognitivo de Cambridge) Roth, Tym, Mountjoy y colaboradores (1986)	Orientación, memoria, lenguaje, construcción (copias), atención, abstracción, percepción, cálculo
Severe Impairment Battery (Batería de deterioro grave) Saxton, McGonigle-Gibson, Swihart y colaboradores (1990)	Orientación, memoria, lenguaje, atención, interacción social, construcción (copias), praxis, percepción visual.

RESUMEN

1. Para propósitos de evaluación, el procesamiento cognitivo se considera como la secuencia que se sigue a través de las siguientes etapas: entrada sensorial, atención y concentración, aprendizaje y memoria, habilidades de lenguaje o habilidades visoespaciales/de manipulación o ambas, funciones ejecutivas y salida motora (figura 9–6).
2. Por lo general, la evaluación de la entrada sensorial se logra a través de la estimulación unilateral y bilateral en las modalidades de tacto, oído y visión. Las tareas típicas (p. ej., localización de dedos) son tan sencillas que las personas sanas pocas veces cometen errores.
3. Las medidas del deterioro en la atención incluyen la resta de sietes en serie; la CPT, una familia de tareas de vigilancia por computadora; y la PASAT, una prueba con límites de tiempo para la aritmética mental (suma de pares sucesivos de dígitos), la cual

es muy sensible a los efectos de la conmoción cerebral.

4. Una respetada prueba de la memoria es la WMS-III, que constituye una revisión sustancial de la escala original publicada hace casi 50 años. Con una estandarización cuidadosa, la WMS-III consta de 17 subpruebas, incluyendo algunas sobre el recuerdo sorpresivo media hora después de la aplicación original.
5. Otra prueba ampliamente utilizada de la memoria es la RAVLT, en la que se lee a la persona evaluada la misma lista de 15 sustantivos concretos en cinco ensayos consecutivos. El recuerdo se somete a prueba después de cada ensayo y también después de la aplicación de una lista interpolada de palabras.
6. La afasia es cualquier desviación del desempeño en el lenguaje, producto de daño cerebral. Las pruebas de la afasia [p. ej., la *Reitan-Indiana Aphasia Screening Test* o el *Boston Diagnostic Aphasia Exam* de Goodglass y Kaplan] evalúan, por lo común, el discurso espontáneo, la repetición de oraciones y frases, la comprensión del lenguaje hablado, la localización de palabras, lectura, escritura, copiado y cálculo).
7. Las pruebas de la capacidad espacial y de manipulación incluyen pruebas de dibujo o copiado como el Test Guestáltico Visomotor y las pruebas de construcción con cubos tridimensionales; ambos tipos son sensibles a los efectos del daño cerebral.
8. Las funciones ejecutivas incluyen el análisis lógico, conceptuación, razonamiento, planificación y flexibilidad de pensamiento. Pruebas útiles para la evaluación de las funciones ejecutivas incluyen el Test de laberintos Porteus; la *Wisconsin Card Sorting Test* y la *Tinkertoy® Test*, llamada así debido a los materiales que utiliza.
9. Las baterías de pruebas neuropsicológicas incluyen por lo común medidas de la salida motora como la Prueba de Golpeteo con los Dedos de la batería Halstead-Reitan. Es típico que la mano no dominante sea 10% más lenta que la mano dominante; las desviaciones con respecto a este patrón pueden indicar una lesión en el hemisferio opuesto al de la mano más lenta.
10. Otras pruebas motoras útiles incluyen el Test no verbal de la Universidad de Purdue, que requiere que la persona examinada coloque clavijas en orificios con la mano izquierda, mano derecha y después con ambas manos, y las tareas simples de trazado de líneas, que llevan a cabo con facilidad la mayoría de los niños de 10 años.
11. La LNNB consta de 269 reactivos separados modelados de acuerdo al trabajo de Luria y estandarizado de manera formal. El autor de la prueba y sus colaboradores informan de excelente confiabilidad y fuerte validez (p. ej., tasas de acierto de 80% o mejores en la identificación de personas con daño cerebral).
12. El Examen del Estado Mental (EEM) es una entrevista definida de manera aproximada que por lo general antecede a otras formas de evaluación psicológica y médica. Las áreas que se evalúan en el EEM incluyen orientación, memoria, pensamiento, sentimiento y juicio.
13. Una prueba de detección del estado mental que es de particular utilidad con ancianos es el MMSE. Esta prueba de 30 reactivos tiene elevada confiabilidad y tasas de acierto en algunas poblaciones de 80 a 90% para la detección de demencia en ancianos.

TÉRMINOS Y CONCEPTOS CLAVE

Conmoción cerebral p. 397
Enfermedad de Alzheimer p. 401
Afasia p. 402
Apraxia p. 403
Funciones ejecutivas p. 406

Ambientes especiales para la evaluación psicológica

CAPÍTULO 10

TEMA 10A
Evaluación escolar

Detección de la preparación para la escuela
Valoración intelectual de niños preescolares
Evaluación de las discapacidades para el aprendizaje y trastornos relacionados
Ejemplo de caso 10–1. Una batería de prueba en la evaluación de la discapacidad para el aprendizaje
Evaluación del TDAH
Evaluación de trastornos emocionales y conductuales
Evaluación de talentos especiales
Resumen
Términos y conceptos clave

En este capítulo se estudian las aplicaciones de las pruebas dentro de dos ambientes característicos: las escuelas y el sistema legal. Aunque muy diversos en muchos sentidos, estos dos ambientes para la evaluación comparten algunos aspectos esenciales. En ambas áreas, las normas legales ejercen una influencia poderosa y restrictiva sobre las prácticas de evaluación. Cuestiones como el diagnóstico y clasificación adecuados son especialmente pertinentes. Por ejemplo, cuando se evalúa la discapacidad para el aprendizaje en un estudiante, el psicólogo escolar observa pautas federales que circunscriben la definición de estos trastornos. De manera similar, cuando se realiza la valoración de un cliente para establecer su competencia para so-

meterse a juicio, el psicólogo forense sigue las normas publicadas que determinan los límites de la capacidad legal. En ambos casos, los preceptos legales moldean, configuran y establecen las prácticas de evaluación.

Seguramente, la evaluación escolar y la forense comprenden dos de los campos más interesantes para las pruebas psicológicas. En el tema 10A, Evaluación escolar, se toma en consideración el empleo de las pruebas psicológicas dentro de las escuelas para propósitos como la detección, evaluación y asignación especial. En el tema 10B, Aplicaciones forenses de la evaluación, se analizan los desafíos únicos que enfrentan los psicólogos forenses que llevan a cabo valoraciones dirigidas a un tribunal. Por supuesto, se revisan y catalogan las pruebas relacionadas. Pero más importante es el hecho de que se centra la atención sobre las cuestiones y retos especiales que se enfrentan dentro de estos medios peculiares.

A pesar de que es cierto que una de las principales aplicaciones de las pruebas escolares es la valoración de las discapacidades para el aprendizaje (que se analizan a profundidad en las siguientes secciones), los psicólogos proporcionan muchos otros servicios de evaluación para las escuelas. Éstos incluyen la detección de la preparación para la escuela, la valoración del déficit de atención y otros problemas conductuales y el examen de inteligencia dotada. En las siguientes secciones se revisan estos temas. En el tema 7A, Pruebas con poblaciones especiales, se describen las prácticas adicionales de evaluación escolar, como la valoración del retraso mental y la evaluación de niños con discapacidades.

DETECCIÓN DE LA PREPARACIÓN PARA LA ESCUELA

El propósito de la **detección** es identificar a niños que están en riesgo, de modo que se les pueda canalizar a una valoración más completa (Kamphaus, 1993), ¿pero "en riesgo" de qué? La respuesta general tiene que ver con la probabilidad de reprobación en los años iniciales de la educación primaria. La noción de riesgo se vincula en gran medida con el concepto de demora en el desarrollo, el cual se refiere a los niños cuyo desarrollo cognitivo está muy por debajo de las expectativas para su edad. Algunos niños identificados con esta etiqueta "se ponen al día" durante el curso de su vida. Para ellos, el término demora del desarrollo es una designación apropiada; seguramente es una etiqueta más optimista y menos afrentosa que la de retraso mental —que con frecuencia es la última consecuencia de la demora en el desarrollo.

Los niños con inteligencia baja están en riesgo sustancial de fracasar en la escuela, lo que explica por qué las pruebas individuales de inteligencia representan un papel importante en la valoración de los niños preescolares, pero requieren de un compromiso sustancial de tiempo (hasta dos horas) y deben aplicarlas profesionales cuidadosamente entrenados; por tanto, y debido a razones de índole práctica, las pruebas individuales de inteligencia no son adecuadas como instrumentos de detección.

El instrumento ideal de detección es una prueba corta que puedan aplicar los maestros, enfermeras escolares y otros individuos que tienen un entrenamiento limitado en cuestiones de evaluación. Además, una prueba sensible de detección es aquella que proporciona una puntuación de corte precisa en la clasificación de los niños como normales o en riesgo. En el contexto de las pruebas de detección pueden ocurrir dos errores. Los niños normales que fallan en la prueba se canalizan como casos de falso positivo (debido a que se les clasifica de manera equivocada como positivos en cuanto a discapacidad potencial). Los niños en riesgo que pasan la prueba se consideran como casos de falso negativo (porque se les clasifica de manera inadecuada como negativos en relación con la discapacidad potencial). El lector debe tener en mente que el propósito de la detección es simplemente la identificación de niños que necesitan una valoración adicional, lo cual significa que los casos de falso positivo recibirán mayor valoración. De aquí que la clasificación errónea de falso positivo conduzca en pocas ocasiones a consecuencias indeseables. No obstante, es

típico que los casos de falso negativo no reciban una valoración posterior, de modo que este tipo de clasificación errónea es potencialmente más grave —debido a que un niño en situación de necesidad se considera como normal. Glascoe (1991) recomienda que un instrumento útil debería producir una tasa de falsos negativos menor a 20% (lo cual implica que se detectará con la prueba a 80% de niños que en realidad están en riesgo) y una tasa todavía menor de falsos positivos, menor de 10% (lo que significa que 90% de los niños normales pasarán la prueba).

Las pruebas de detección son breves por definición y, por tanto, están propensas a errores de medición. Incluso con las mejores pruebas disponibles, algunos niños que merecen atención pasarán desapercibidos (casos de falso negativo) y no se reconocerá que necesitan intervención hasta mucho después de haber iniciado la educación primaria. Sin embargo, si una prueba de detección identifica una alta proporción de casos positivos verdaderos —preescolares clasificados en riesgo, quienes necesitan realmente los servicios especiales o una demora en el ingreso a la escuela— dicha prueba satisface un propósito útil.

INSTRUMENTOS PARA DETECCIÓN PREESCOLAR

Aunque se han producido docenas de instrumentos para detectar las demoras en el desarrollo (Malcolm, 1998), aquí se limita el análisis a algunas de ellas: los *Developmental Indicators for the Assessment of Learning –Revised*, DIAL-R (Indicadores del Desarrollo para la Evaluación del Aprendizaje Revisados), la Denver II, una revisión de la *Denver Developmental Screening Test –Revised* (Prueba Denver de Detección del Desarrollo –Revisada), y la *Home Observation for the Measurement of the Enviroment*, HOME (Observación en el Hogar para la Medición del Ambiente). Las primeras dos utilizan enfoques convencionales para la identificación de la demora en el desarrollo, mientras el tercer instrumento, el HOME, engloba una desviación radical con respecto a los procedimientos tradicionales.

DIAL-R

Los *Developmental Indicators for the Assessment of Learning –Revised*, DIAL-R (Indicadores del Desarrollo para la Evaluación del Aprendizaje Revisados) son un procedimiento de detección que se aplica de manera individual y que se diseñó para la detección rápida y eficaz de los problemas del desarrollo (o la inteligencia dotada) en niños preescolares de 2-0 a 5-11 años de edad (Mardell-Czudnowski y Goldenberg, 1990). La prueba analiza el desempeño de los niños en tres dominios del desarrollo: motricidad, conceptos y lenguaje. Cada uno se evalúa a través de ocho reactivos que se subdividen en tareas separadas. Los DIAL-R también proporcionan una evaluación por observación de los problemas socioemocionales (p. ej., llanto o gimoteo, distracción, resistencia, indisciplina), pero de manera menos rigurosa. Los siguientes son ejemplos de reactivos de la prueba dentro de los tres dominios del desarrollo:

1. **Motricidad**: cachar, saltar, construir, cortar, copiar formas y letras, escribir el propio nombre.
2. **Conceptos**: identificación de partes del cuerpo, nombrar colores, contar cubos, palabras de posición, nombrar letras, clasificar.
3. **Lenguaje**: decir palabras individuales, dar sus datos personales, decir sustantivos, clasificar alimentos, solucionar problemas orales, extensión de oraciones habladas.

La calificación para algunos reactivos es separada y objetiva, mientras que para otras preguntas los criterios de calificación del manual dejan oportunidad para la interpretación subjetiva, lo cual disminuye la confiabilidad del instrumento. La calificación total se obtiene de la suma de las tres puntuaciones de área. Para cada puntuación de área y también para la calificación total, el manual proporciona puntuaciones de corte para asignar al niño a 1 de 3 grupos de resultado denominados "aprobado", "problema potencial" o "potencialmente avanzado". Las puntuaciones de corte de los DIAL-R están disponibles para tres grupos normativos:

caucásico, minoritario y censal (grupos combinados). Las muestras de estandarización son muy similares a los datos del censo de 1990 en cuanto a antecedentes educativos de los padres.

La confiabilidad de los DIAL-R es adecuada, en vista de que es una prueba breve con propósitos de detección. Los datos de confiabilidad *test-retest* ($N = 65$) indican coeficientes promedio que van de .76 para el dominio de motricidad al .90 para conceptos. La validez del instrumento se ha estimado en las pautas familiares de contenido, constructo y relacionada con el criterio. La validez de contenido se considera alta y se obtuvo a través de un panel de ocho expertos que proporcionaron revisiones del contenido y ayudaron a eliminar los reactivos inapropiados o tendenciosos. La validez de constructo es más cuestionable debido a que la evidencia del análisis factorial no parece sustentar la validez de los tres dominios como subpruebas independientes.

Es con respecto a la validez relacionada con el criterio que los DIAL-R han hecho surgir el mayor escepticismo. El valor de una prueba de detección puede juzgarse mejor en función del grado al que identifica **con precisión** a los niños que necesitan mayor evaluación relativa al desarrollo. Una prueba con buena validez relacionada con el criterio clasifica de manera correcta a una alta proporción de niños que en la actualidad obtienen puntuaciones bajas en pruebas más amplias de demora del desarrollo (validez concurrente) o que posteriormente se confirma que presentan dichos problemas (validez predictiva). Un estadístico útil para estimar la validez relacionada con el criterio es la **sensibilidad**, que es la proporción de casos problema confirmados que se "señalaron" de manera precisa como tales. Por desgracia, los DIAL-R no constituyen una prueba eficaz en este sentido. El manual cita una variedad de estudios que revelan sensibilidad débil, en general dentro del rango de 25 a 57%. Lo que esto significa es que de aquellos niños que en realidad tienen un problema, sólo se identificó de 25 a 57% como "problema potencial" con los DIAL-R. La prueba funciona mejor en la identificación correcta de niños normales como tales, con una **especificidad** de cerca de 90%. No obstante, la baja sensibilidad implica que muchos niños en necesidad de consideración especial pasarán desapercibidos para esta prueba (Cooper y Shepard, 1992).

DENVER II

La Denver II (Frankenburg, Dodds, Archer y col., 1990) es una versión actualizada de la muy popular *Denver Developmental Screening Test –Revised* (Prueba Denver de Detección del Desarrollo –Revisada; Frankenburg, 1985; Frankenburg y Dodds, 1967). Es probable que ésta sea la herramienta de detección pediátrica más conocida e investigada en EUA. El instrumento es popular a nivel mundial —se le ha traducido a 44 idiomas diferentes. Adecuada para lactantes y niños de 1 mes a 6 años de edad, la prueba consta de 125 reactivos en cuatro áreas: personal-social, motora fina-adaptativa, lenguaje y motora gruesa. Los reactivos son una mezcla de informes de los padres, obtención directa de conductas y observación. Cada uno está dispuesto en orden cronológico en la prueba según la edad del niño, y se califica con aprobado/reprobado. La prueba comienza a un nivel apropiado para la edad y continúa hasta que el niño falla en tres reactivos. El tiempo total de valoración es de 20 minutos o menos.

A diferencia de otras pruebas de detección, la Denver II no produce un cociente o puntuación del desarrollo. En vez de ello, los resultados en cerca de 30 reactivos apropiados para la edad proporcionan una puntuación que se puede interpretar como normal, cuestionable o anormal, con referencia a normas basadas en la edad. También se incluye una categoría de "inestable". La muestra de estandarización se aplicó a 2 096 niños, todos del estado de Colorado, estratificados conforme a edad, raza y nivel socioeconómico.

En reconocimiento a que ningún instrumento es perfectamente confiable ya que los niños cambian a lo largo del tiempo, los autores de la Denver II recomiendan la repetición de la prueba a intervalos de aproximadamente seis meses

hasta los dos años de edad, y después cada año hasta los cinco años. Se informa que la Denver II tiene una confiabilidad sobresaliente para una prueba breve de detección. La confiabilidad intercalificadores entre calificadores entrenados tuvo un promedio notable de .99. La confiabilidad *test-retest* para la calificación total en un intervalo de 7 a 10 días promedió .90.

La prueba Denver posee excelente validez de contenido en cuanto a que las conductas examinadas tienen el reconocimiento de las autoridades en desarrollo infantil como indicadores importantes del desarrollo. No obstante, las categorías de interpretación de la prueba (normal, cuestionable, anormal) se basaron en el juicio clínico y, por tanto, esperan un estudio adicional para validación. Algunos estudios iniciales hacen surgir preocupaciones importantes. Glascoe y Byrne (1993) valoraron a 89 niños en ambientes de cuidado diurno que tenían de 7 a 70 meses de edad. Con base en una amplia valoración independiente, en 18 de ellos se confirmó la presencia de demoras del desarrollo de acuerdo con las definiciones federales para trastornos discapacitantes (p. ej., demoras en el lenguaje, retraso mental y autismo). Aunque la Denver II funcionó bien en la identificación correcta de 15 de los 18 niños en riesgo, el instrumento tuvo un desempeño deficiente con los niños normales. De hecho, 38 de los 71 niños normales fallaron en la prueba y se les clasificó como cuestionables o anormales. En general, casi 4 de cada 6 niños que realizan la prueba se canalizarían para evaluación adicional, y de los 4, sólo 1 tendría una verdadera discapacidad. Los investigadores concluyen que:

> Esto causa ansiedad y gastos innecesarios para los padres, desperdicia preciosos recursos diagnósticos y de intervención, y deja a los profesionales con preguntas sin responder acerca de la condición de desarrollo de los niños, tanto antes como después de la detección.

Estos autores recomiendan mayores estudios de validación dando la importancia adecuada y eliminando en lo posible algunos reactivos antes de que la prueba pueda utilizarse de manera general.

HOME

La *Home Observation for Measurement of the Environment*, HOME (Observación en el Hogar para la Medición del Ambiente), conocida popularmente como Inventario HOME, es con toda probabilidad el índice más utilizado del ambiente de los niños. Basada en la observación dentro del hogar y con una entrevista con la principal persona encargada del cuidado del niño, el instrumento proporciona una medición de los ambientes físico y social de éste. El Inventario HOME viene en tres formas: Lactantes e Infantes, Primera Infancia e Infancia Media. Las últimas ediciones del instrumento, con fecha de 1984, surgieron después de 15 años de revisión y refinamiento metódicos (Caldwell y Richmond, 1967; Caldwell y Bradley, 1984, 1994).

ANTECEDENTES Y DESCRIPCIÓN

Antes del desarrollo del Inventario HOME, la medición de los ambientes de los niños se basaba, en gran medida, en datos demográficos como educación, ocupación, ingresos y lugar de residencia de los padres. Con frecuencia, estos índices se combinaban en una medida acumulativa conocida como clase social o nivel socioeconómico (NSE). Por ejemplo, Hollingshead y Redlich (1958) desarrollaron un continuo de clase social derivado de la residencia, ocupación y educación del jefe de familia. La puntuación de NSE de una familia cuyo jefe trabajara en un empleo de oficina, tuviera diploma de educación media y viviera en un área residencial de nivel medio, se calcularía de la siguiente manera (Hollingshead y Redlich, 1958):

Factor	Valor de Escala	x	Ponderación del factor	=	Puntuación parcial
Residencia	3		6		18
Ocupación	4		9		36
Educación	4		5		20
Indice de NSE				=	74

Para propósitos de investigación, los científicos sociales pueden categorizar a las familias en una jerarquía de cinco clases sociales (clases I a V) con base en la calificación total. El lector observará que la medición de Hollingshead y Redlich se obtuvo completamente de los índices de **estatus**. La suposición no expresada es que estos índices reflejan, de manera indirecta, una variación ambiental significativa. Dicho en términos directos, los defensores del NSE como medida ambiental creen que, en promedio, los niños de una clase social más alta tendrán un ambiente mejor que aquéllos de una más baja.

En contraste con el enfoque del NSE, el Índice HOME se desarrolló para proporcionar una medida directa de **proceso** de los ambientes infantiles. La filosofía guía de este instrumento es que la evaluación directa de las experiencias de los niños es un mejor índice del ambiente en el hogar que medidas indirectas como la ocupación y educación de los padres. Aunque es cierto que la clase social —como lo representa la ocupación, educación y residencia— proporciona una medida oblicua de la riqueza del ambiente, los autores del Inventario HOME discutirían que la evaluación directa de las experiencias de los niños proporciona un índice más preciso de las variaciones en el ambiente hogareño. Así, la evaluación con este inventario implica, en parte, la observación directa de los ambientes hogareños de los niños para determinar si ciertos tipos de interacciones y experiencias están presentes o ausentes. Por ejemplo, durante una visita de una hora, el examinador observa si el padre o la madre se comunican espontáneamente con el niño cuando menos en cinco ocasiones, determina si éste tiene cuando menos 10 libros o discos con cuentos y evalúa si el vecindario es estéticamente agradable según normas detalladas, por citar sólo algunos ejemplos.

El propósito del Inventario HOME es medir la calidad y cantidad de estimulación y apoyo para el desarrollo cognitivo, social y emocional que se encuentra disponible para el niño en su hogar. Las escalas y reactivos del HOME se obtuvieron de una lista de procesos ambientales identificados en la investigación y teoría considerados importantes para el desarrollo óptimo en la infancia (Caldwell y Bradley, 1984). Estos procesos promotores del crecimiento incluyen la satisfacción de necesidades básicas; contacto frecuente con un número relativamente pequeño de adultos; un clima emocional positivo que fomente la confianza en sí mismo y en los demás; información sensorial apropiada, variada y con un cierto patrón; consistencia en las respuestas físicas, verbales y emocionales de los demás; un mínimo de restricciones sociales sobre la conducta de exploración y motora; estructura y orden en el ambiente cotidiano; dar e interpretar experiencias culturales variadas de adultos; materiales y ambientes apropiados de juego; contacto con adultos que valoren los logros y programación acumulativa de experiencias que estén al nivel de desarrollo del niño (Caldwell y Bradley, 1984). Por tanto, en pocas palabras, el propósito del HOME consiste en medir patrones específicos y designados de nutrición y estimulación disponibles para los niños en el hogar.

Para llenar el Inventario HOME, el examinador debe observar al niño y a la persona encargada de su cuidado (por lo general la madre) en interacción dentro del ambiente del hogar. Las calificaciones de algunos de los reactivos del inventario se logran de la observación del ambiente físico. Además, la respuesta para algunos de ellos se basa en autoinformes de la persona encargada del niño. Los reactivos se califican de manera dicotómica, 1 para presencia, 0 para ausencia. Por ejemplo, un reactivo pregunta si se incluye al niño en la compra de abarrotes al menos una vez por semana. El manual del inventario alienta un enfoque relajado, semiestructurado, para la observación y entrevista (Caldwell y Bradley, 1984). Completar el inventario requiere cerca de una hora.

El HOME está disponible en tres formas: Lactantes e Infantes (0 a 3 años), Infancia Temprana (3 a 6 años) e Infancia Media (6 a 10 años). La forma para Lactantes e Infantes consta de 45 reactivos organizados en las siguientes seis subescalas:

Respuesta emocional y verbal del padre.
Aceptación de la conducta del niño.
Organización del ambiente.

Suministro de materiales apropiados de juego.
Participación del padre o madre con el niño.
Variedad de estimulación.

La versión para Infancia Temprana tiene 55 reactivos organizados en ocho subescalas, e Infancia Media 59 reactivos distribuidos en ocho subescalas. En el cuadro 10–1 se listan los reactivos para la versión del Inventario HOME dirigida a Lactantes e Infantes. En Caldwell y Bradley (1984) se pueden encontrar detalles acerca de los reactivos específicos que se incluyen en el HOME.

CARACTERÍSTICAS TÉCNICAS

Las normas relacionadas con el Inventario HOME están disponibles en varias fuentes. Para la versión de Lactantes e Infantes, Caldwell y Bradley (1984) informan de medias y desviaciones estándar por subescala para 174 familias de Little Rock, Arkansas. En comparación con la población general, esta muestra parece sobrerrepresentar a las familias de NSE bajo; por ejemplo, 34% de ellas estaban en la beneficencia del estado, y 29% eran hogares de un solo padre. Para la versión de Infancia Temprana, los datos de estandarización se obtuvieron con 232 familias en Little Rock, con una representación excesiva similar de familias de NSE bajo. En el caso de la versión para Infancia Media, Bradley y Rock (1985) informaron de medias y desviaciones estándar para 141 familias de Little Rock. Aproximadamente la mitad de éstas eran afroestadounidenses y las restantes caucásicas; niños y niñas tuvieron una representación igual en la muestra. Se consideró que estas familias eran representativas de todas aquellas que crían a niños en edad escolar en Little Rock, Arkansas; no obstante, en las tres versiones es evidente que las muestras de estandarización proporcionan sólo normas locales. Estos datos pueden ser útiles como puntos de referencia, pero no deben igualarse con una muestra nacional, aleatoria y estratificada.

La confiabilidad del Inventario HOME se ha demostrado de diferentes maneras, en particular para la versión de Lactantes e Infantes, que es la que se analiza aquí. Los autores señalan que los estudios *test-retest* a corto plazo son inapropiados, dado que es bastante probable que la persona que responde al inventario recuerde una respuesta específica dada a una pregunta, lo cual inflaría de manera artificial las correlaciones *test-retest* (Bradley y Caldwell, 1984). Los métodos utilizados para la evaluación de la confiabilidad incluyeron concordancia interobservadores, consistencia interna y coeficientes de estabilidad *test-retest* de amplio rango para 91 familias de la muestra de estandarización. Por definición, se informa que la concordancia interobservadores para los reactivos de subescala es de 90% o mayor, dado que éste es el criterio de entrenamiento para los nuevos calificadores. Los estimados de consistencia interna con la fórmula Kuder-Richardson 20 van de .67 a .89 para todas las subescalas, excepto Variedad de estimulación, la cual produjo un coeficiente de sólo .44, que se debió al pequeño número de reactivos en la subescala (cinco). Los datos *test-retest* se obtuvieron con 91 familias evaluadas cuando su lactante/infante tenía 6, 12 y 24 meses de edad. Los coeficientes indicaron un grado de moderado a alto de estabilidad para las subescalas, donde la mayor parte de las correlaciones se colocaron en los rangos de .50, .60 y .70. La correlación entre la puntuación total de la prueba a los 12 y 24 meses de edad llegó a un respetable .77.

La validez del Inventario HOME se ha consolidado a través de datos de investigación que muestran correlaciones modestas con los índices de NSE. Debido a que el inventario se propuso como un índice más significativo y sensible del ambiente que la clase social, las puntuaciones del HOME deberían relacionarse de manera significativa, pero no elevada, con los índices de NSE. En el caso de la versión para Lactantes e Infantes, las correlaciones entre las subescalas del Inventario HOME y el NSE se encuentran principalmente en los rangos de .30 y .40, mientras que aquélla entre la puntuación total y el NSE es de .45 (Bradley, Rock, Caldwell y Brisby, 1989).

Las calificaciones del HOME también muestran relaciones fuertes y confirmatorias

Cuadro 10–1. Lista de subescalas y reactivos para la versión de Lactantes e infantes del Inventario HOME

Inventario HOME

Coloque un signo de más (+) o menos (–) en la casilla junto a cada reactivo si la conducta se observó durante la visita o si el padre o madre informa que las condiciones o hechos son característicos del ambiente en el hogar. Anote el subtotal y el total en la cara frontal de la hoja de registro.

I. RECEPTIVIDAD emocional y verbal

☐ 1. El progenitor vocaliza de manera espontánea hacia el niño dos veces

☐ 2. El progenitor responde verbalmente a las vocalizaciones del niño

☐ 3. El progenitor dice al niño el nombre del objeto o persona durante la visita

☐ 4. El habla del progenitor es precisa y audible

☐ 5. El progenitor inicia intercambios verbales con el visitante

☐ 6. El progenitor conversa libremente y con facilidad

☐ 7. El progenitor permite que el niño participe en juego "desordenado"

☐ 8. El progenitor alaba en forma espontánea al niño al menos en dos ocasiones

☐ 9. La voz del progenitor transmite sentimientos positivos hacia el niño

☐ 10. El progenitor acaricia o besa al niño al menos una vez

☐ 11. El progenitor responde de manera positiva a la alabanza del visitante hacia el niño

☐ Subtotal

II. ACEPTACIÓN de la conducta del niño

☐ 12. El progenitor no le grita al niño

☐ 13. El progenitor no expresa molestia u hostilidad hacia el niño

☐ 14. El progenitor no golpea al niño durante la visita

☐ 15. No más de un caso de castigo físico durante la última semana

☐ 16. El progenitor no regaña o critica al niño durante la visita

☐ 17. El progenitor no interfiere o restringe al niño en más de tres ocasiones

☐ 18. Están presentes y visibles cuando menos 10 libros

☐ 19. La familia tiene una mascota

☐ Subtotal

III: ORGANIZACIÓN del ambiente

☐ 20. Uno de tres sustitutos regulares proporciona cuidado alterno

☐ 21. Se lleva al niño a la tienda de abarrotes cuando menos una vez por semana

☐ 22. El niño sale de casa cuando menos cuatro veces por semana

☐ 23. Al niño se le lleva con regularidad al consultorio del médico o a la clínica

☐ 24. El niño tiene un lugar especial para sus juguetes y tesoros

☐ 25. El ambiente de juego del niño es seguro

☐ Subtotal

IV. Proporcionar MATERIALES DE JUEGO

☐ 26. Juguetes o equipo para actividad muscular

☐ 27. Juguete para empujar o jalar

☐ 28. Carreola o andadera, carrito, patineta o triciclo

☐ 29. El progenitor le proporciona juguetes al niño durante la visita

☐ 30. Equipo de aprendizaje apropiado para la edad —muñecos suaves o para *role-playing*

☐ 31. Facilitadores del aprendizaje —móvil, mesa y sillas, silla alta, corralito

☐ 32. Juguetes para coordinación simple ojo-mano

☐ 33. Juguetes para coordinación compleja ojo-mano (aquellos que permiten combinaciones)

☐ 34. Juguetes para literatura o música

☐ Subtotal

Cuadro 10–1. Lista de subescalas y reactivos para la versión de Lactantes e infantes del Inventario HOME (continuación)

V. PARTICIPACIÓN de los padres con el niño

☐ 35. El progenitor mantiene al niño dentro de su rango de visión, lo mira con frecuencia

☐ 36. El progenitor le habla al niño mientras realiza las tareas del hogar

☐ 37. El progenitor alienta conscientemente el progreso del desarrollo

☐ 38. El progenitor inventa juguetes que tienen valor a través de la atención personal

☐ 39. El progenitor estructura los periodos de juego del niño

☐ 40. El progenitor le proporciona al niño juguetes que lo desafían a desarrollar nuevas habilidades

☐ Subtotal

VI. Oportunidades de VARIEDAD

☐ 41. El padre proporciona cierto cuidado diariamente

☐ 42. El progenitor lee cuentos al niño al menos tres veces por semana

☐ 43. El niño toma cuando menos una de sus comidas diarias con la madre o el padre

☐ 44. La familia visita a familiares o recibe visitas una vez al mes aproximadamente

☐ 45. El niño tiene tres o más libros propios

☐ Subtotal

☐ PUNTUACIÓN TOTAL

Para el fraseo completo de los reactivos, por favor refiérase al Manual de aplicación.

Fuente: Reproducido con autorización de Caldwell, B. M. y Bradley, R. H. (1984). *Home Observation for Measurement of the Environment.* Little Rock: AR: University of Arkansas at Little Rock.

de la teoría con criterios externos apropiados que incluyen desarrollo del lenguaje y cognitivo, fracaso escolar, intervención terapéutica y retraso mental (Caldwell y Bradley, 1984). Las correlaciones entre las puntuaciones del HOME y medidas intelectuales como las *Bayley Scales of Infant Development* (Escalas de Desarrollo Infantil Bayley) y la Stanford-Binet son particularmente informativas. Bradley y Caldwell (1984) llevaron a cabo un estudio longitudinal con 174 familias, en el cual aplicaron el HOME a los 6, 12 y 24 meses de edad y correlacionaron estas calificaciones con el Índice de Desarrollo Mental de Bayley (MDI) a los 12 meses y con el CI de la Stanford-Binet a los 36 y 54 meses de edad. El patrón de correlaciones indicó una validez predictiva más alta que la concurrente, es decir, las correlaciones entre el HOME y el CI a los 36 y 54 meses de edad fueron mayores que aquéllas entre el HOME y el MDI a los 12 meses de edad (cuadro 10–2). Los estudios analítico factoriales del HOME también sustentan la validez de constructo de este instrumento (Bradley, Mundfrom, Whiteside y colaboradores, 1994; Mundfrom, Bradley y Whiteside, 1993).

Además de su utilidad como herramienta de investigación, el HOME representa una promesa como instrumento clínico. Debido a que las calificaciones bajas en el HOME predicen riesgo de discapacidad intelectual, el inventario puede utilizarse para identificar a niños para quienes sería apropiada la intervención terapéutica. Este tipo de intervención es más que sólo una aplicación humanista de los descubrimientos de investigación —puede ser un requisito de ley:

> Entre las nuevas razones más apremiantes para investigar el ambiente de los niños con discapacidades es el requisito en la ley pública 99-457 de que todos los planes de servicio para los niños en edad escolar que tienen discapacidades incluyan un componente para los padres. Ya no será suficiente con tener un plan de regularización

Cuadro 10–2. Correlaciones entre las puntuaciones del HOME a los 12 meses y puntuaciones cognitivas a los 12, 36 y 54 meses

Subescala del HOME	MDI a los 12 meses	CI a los 36 meses	CI a los 54 meses
Receptividad	.15	.39*	.34*
Restricción	.01	.24*	.21
Organización	.20	.39*	.34*
Materiales de juego	.28*	.56*	.52*
Participación	.28*	.47*	.36*
Variedad	.05	.28*	.32*
Puntuación total	.30*	.58*	.53*

*$p < .05$

Nota: MDI se refiere al Índice de Desarrollo Mental de las *Bayley Scales of Infant Development* (Escalas Bayley de Desarrollo Infantil); el CI es de la Stanford-Binet.

Fuente: Reproducido con autorización de Bradley, R. H. y Caldwell, B. M. (1984). 174 children: A study of the relationship between home environment and cognitive development during the first 5 years. En A.W. Gottfried (ed.) *Home environment and early cognitive development: Longitudinal Research.* Orlando, FL: Academic Press.

dirigido en exclusiva al niño. También deben incluirse planes para la participación y desarrollo de las capacidades de los padres (Bradley, Rock, Caldwell y Brisby, 1989).

En resumen, el Inventario HOME es prometedor no sólo en cuanto a investigación, sino también como complemento práctico para la intervención.

VALORACIÓN INTELECTUAL DE NIÑOS PREESCOLARES

Por definición, las pruebas de detección son menos precisas que las evaluaciones completas, es decir, tan sólo señalan la necesidad de mayor valoración. Como parte de cualquier valoración de seguimiento, un profesional puede aplicar una variedad de instrumentos que casi con toda seguridad incluirán una prueba individual de inteligencia. Aunque las puntuaciones de prueba de los preescolares en medidas de inteligencia son notablemente inestables a la larga, es probable que a corto plazo no exista un mejor índice de la probabilidad de que un niño esté en riesgo de fallar en la escuela.

Varias pruebas de inteligencia de aplicación individual son adecuadas para niños preescolares. Las de uso más amplio son la Stanford-Binet: Cuarta Edición (SB:CE), la Escala de Inteligencia para los Niveles Preescolar y Primario (WPPSI-R), la Batería de Evaluación de Kaufman para Niños (K-ABC) y las *Differential Ability Scales*, DAS (Escalas de Capacidad Diferencial). Estos instrumentos se revisaron en un tema anterior.

EVALUACIÓN DE DISCAPACIDADES PARA EL APRENDIZAJE Y TRASTORNOS RELACIONADOS

El campo de las discapacidades para el aprendizaje (DA) es una de las áreas de crecimiento más rápido dentro de la evaluación; paradójicamente, también es uno de los campos más polémicos y complejos de las pruebas psicológicas. Se requiere de antecedentes considerables para comprender la función de las pruebas psicológicas en la valoración de las discapacidades para el aprendizaje. Se comenzará con una pregunta aparentemente sencilla que resulta tener una respuesta complicada: ¿qué es una **discapacidad para el aprendizaje?**

DEFINICIÓN FEDERAL DE LAS DISCAPACIDADES PARA EL APRENDIZAJE

Durante decenios, la naturaleza esencial de las discapacidades se ha comprendido en términos de la definición incluida en la ley federal. En 1975, el Congreso de Estados Unidos aprobó

la Ley Pública 94-142, el Decreto de Educación para todos los Niños con Discapacidad. Una de las medidas de este decreto era la siguiente definición de las discapacidades para el aprendizaje:

> El término "discapacidad específica para el aprendizaje" significa un trastorno en uno o más de los procesos psicológicos básicos implicados en la comprensión o utilización del lenguaje, hablado o escrito, que se puede manifestar en una capacidad imperfecta para escuchar, hablar, leer, escribir, deletrear o llevar a cabo cálculos matemáticos. El término incluye padecimientos como impedimentos perceptuales, lesión cerebral, disfunción cerebral mínima, dislexia y afasia del desarrollo. El término no incluye a niños que tienen discapacidades para el aprendizaje que son principalmente resultado de impedimentos visuales, auditivos o motores; retraso mental, trastorno emocional, o desventajas ambientales, culturales o económicas (USDE, 1977, p. 65083).

El compromiso con una definición decretada por el Estado se confirmó en 1990 con la aprobación de la Ley Pública 101-476, el Decreto de Educación para Individuos con Discapacidades (IDEA). Un poco más de la mitad de los estados en EUA siguen ahora este modelo. Los estados restantes decretan enfoques similares (Lerner, 1993).

La definición federal que se incorporó en el IDEA también estipula un enfoque operacional para la identificación de niños con discapacidades para el aprendizaje. En términos específicos, los candidatos para un diagnóstico de DA deben demostrar una discrepancia grave entre su capacidad general (inteligencia) y el aprovechamiento específico en una o más de estas siete áreas:

Expresión oral.
Comprensión auditiva.
Expresión escrita.
Habilidad básica para la lectura.
Comprensión de lectura.
Cálculo matemático.
Razonamiento matemático.

El modelo de discrepancia para la identificación de niños con DA ha funcionado como un indicador para los psicólogos escolares. En efecto, el modelo decreta que los psicólogos deben aplicar una prueba individual de inteligencia (medida de capacidad general) y una de rendimiento (medida de rendimiento específico) y después buscar una discrepancia entre el CI Total y una o más de las áreas de rendimiento escolar (p. ej., lectura, matemáticas, expresión escrita).

En términos prácticos, la discrepancia grave se ha definido como una diferencia de una desviación estándar o más entre la inteligencia general y el rendimiento específico. Una práctica común en la identificación de niños con DA consiste en comparar el CI Total en una prueba individual de inteligencia como el WISC-III con las calificaciones de rendimiento específico en una prueba individual de rendimiento como el WIAT (*Wechsler Individual Achievement Test*) o un instrumento similar que tenga subpruebas cuyas normas tengan una media de 100 y desviación estándar de 15. Una diferencia de 15 puntos o más entre el CI Total y el rendimiento específico en cualesquiera de las áreas antes mencionadas daría lugar a la sospecha de una discapacidad para el aprendizaje.

Por desgracia, la definición federal no ha cumplido con los propósitos deseados y cada vez es más frecuente que los psicólogos escolares y otros profesionales se acojan a otros enfoques para comprender y evaluar las discapacidades para el aprendizaje en los niños. El problema fundamental es que muchos, pero muchos niños que exhiben graves problemas de aprendizaje en la escuela y se beneficiarían de los servicios dirigidos a la DA, no satisfacen los criterios psicométricos de una discrepancia grave. Esto es debido a que una discapacidad para el aprendizaje puede afectar de manera adversa el desempeño **tanto** en la prueba de inteligencia **como** en las medidas de rendimiento que se utilizan para diagnosticar el trastorno, lo cual da por resultado un perfil de prueba que no encaja con el modelo de discrepancia pero que, no obstante, representa una DA (Shaw, Cullen, McGuire y Brinckerhoff, 1995). Otro problema es que cada estado ha adoptado diferentes fórmulas de discrepancia, de modo que

el diagnóstico de un niño puede ser de DA en un lugar, pero no en otro. Esto no sólo crea confusión, sino que socava la integridad de toda la empresa de identificación de las DA.

SURGIMIENTO DEL CONSENSO: UNA NUEVA DEFINICIÓN DE DISCAPACIDADES PARA EL APRENDIZAJE

Después de un largo periodo de confusión y lucha acerca de la definición de las discapacidades para el aprendizaje, los especialistas y educadores comenzaron a coincidir alrededor de una perspectiva consensual a inicios del decenio de 1990. El *National Joint Committee on Learning Disabilities*, NJCLD (Comité Nacional Conjunto sobre Discapacidades para el Aprendizaje), un grupo de representantes de ocho organizaciones nacionales con un interés especial en las discapacidades para el aprendizaje, propuso una nueva definición. Aunque similar a la definición federal, el nuevo enfoque contiene contrastes importantes:

> **Discapacidades para el aprendizaje** es un término general que se refiere a un grupo heterogéneo de trastornos que se manifiestan a través de dificultades significativas para la adquisición y utilización de las capacidades de atención auditiva, habla, lectura, escritura, razonamiento o matemáticas. Estos trastornos son intrínsecos del individuo, se supone que se deben a una disfunción del sistema nervioso central y pueden ocurrir a lo largo de la vida. Pueden existir problemas en conductas de autorregulación, percepción social e interacción social junto con las discapacidades para el aprendizaje, pero en sí mismos no constituyen una discapacidad para el aprendizaje. Aunque éstas pueden ocurrir junto con otros padecimientos incapacitantes [p. ej., deterioro sensorial, retraso mental (RM), grave trastorno emocional (TE)] o con influencias extrínsecas (como diferencias culturales, instrucción insuficiente o inapropiada), no son el resultado de tales padecimientos o influencias (NJCLD, 1988, p. 1).

La nueva definición evita la referencia vaga a "procesos psicológicos básicos", especifica que el trastorno es intrínseco del individuo, identifica la disfunción del sistema nervioso central como origen de los problemas de DA y expresa de manera explícita que éstos pueden persistir hasta la edad adulta.

Quizá lo más importante de todo es que el enfoque del NJCLD abandona la dependencia excesiva en la discrepancia entre capacidad y rendimiento como sello distintivo de la DA. En vez de ello, el nuevo modelo especifica que la condición necesaria (pero no suficiente) para la DA es que el individuo (niño o adulto) exhiba una debilidad intraindividual en una o más de las áreas centrales del funcionamiento académico (capacidades de atención auditiva, habla, lectura, escritura, razonamiento o matemáticas). Shaw *et al.* (1995) ilustran cómo podría verse el modelo del NJCLD dentro de la práctica (figura 10–1). En este enfoque, la primera tarea consiste en identificar una o más debilidades intraindividuales como áreas centrales. Éstas siempre son relativas a las fortalezas en varias otras áreas centrales. En otras palabras, las personas que tienen lento aprendizaje en todas las áreas no satisfacen los criterios de DA. El segundo paso consiste en rastrear las dificultades para el aprendizaje hasta el sistema nervioso central, las cuales pueden manifestarse como problemas con el procesamiento de información. Por ejemplo, una adulta joven con una grave debilidad en atención auditiva (como puede juzgarse por su discapacidad para aprender con el enfoque tradicional de enseñanza por medio de disertación) podría exhibir un déficit en una prueba de memoria verbal —lo cual confirmaría que en el centro de su discapacidad se encuentra un problema con el procesamiento de información. El propósito del tercer paso (examen de habilidades psicosociales y de capacidades físicas y sensoriales) consiste en especificar los problemas adicionales que pueden requerir de atención para propósitos de planificación de un programa. Por último, en el cuarto paso, el examinador descarta explicaciones ajenas a la DA para las dificultades de aprendizaje (dado que éstas obligarían a tomar una estrategia diferente para la regularización).

Hammill (1990) ha señalado sabiamente que las realidades políticas son tales que es

Paso 1. Discrepancia intraindividual

El examinador identifica una dificultad significativa en una o más áreas centrales junto con fortalezas relativas en varias áreas. Áreas Centrales: Audición, Habla, Lectura, Escritura, Razonamiento, Matemáticas, Áreas temáticas.

Paso 2. Discrepancia intrínseca del individuo

El examinador rastrea el origen de la discrepancia hasta una disfunción del Sistema Nervioso Central (p. ej., lesión cerebral) o vincula la discrepancia con problemas en el procesamiento de información (p. ej., memoria, organización o eficiencia del aprendizaje).

Paso 3. Consideraciones relacionadas

El examinador valora la relevancia de las habilidades psicosociales, físicas y sensoriales para la discapacidad para el aprendizaje.

Paso 4. Explicaciones alternativas

El examinador descarta explicaciones alternativas; por ejemplo, factores ambientales, culturales o económicos; o instrucción inapropiada o inadecuada.

Paso 5. Diagnóstico de DA

El examinador determina que los niños que presentan las características de los pasos 1 a 4 satisfacen los criterios para un diagnóstico de DA.

Figura 10–1. Expresión Operacional de la Definición de Discapacidades para el Aprendizaje del NJCLD.
Fuente: Basado en Brinckerhoff, L., Shaw, S. y McGuire, J. (1993). *Promoting Postsecondary Education for Students with Learning Disabilities: A Handbook for Practitioners.* Austin, TX: PRO-ED.

posible que la definición del NJCLD nunca reemplace al enfoque decretado por el estado. Pero esto es menos importante que el hecho de que los padres y profesionales se unan alrededor de una definición con propósitos de investigación y comunicación entre sí. En la actualidad, la definición del NJCLD ha recibido el más fuerte apoyo general de los profesionales en el campo de la evaluación.

Aunque no se analizarán aquí los puntos de vista adicionales, se puede referir al lector a varios otros modelos destacados de las discapacidades para el aprendizaje (*Special Education Today,* 1985; Lerner, 1993; Lyon, 1994). En este contexto, y aunque se ofreció hace varios años, vale la pena mencionar el sabio consejo de Farnham-Diggory (1978). Poco después de la aprobación de la Ley Pública 94-142 en 1977 con su influyente definición de las discapacidades para el aprendizaje, esta autora escribió:

> La publicación de esta definición ha equivalido a ondear una bandera roja frente de una manada de toros —padres y profesionales por igual. Lejos de aclarar la situación, la definición inspiró tantos gruñidos y coces que el polvo conceptual se ha elevado más espeso que nunca. Parte del problema surge del hecho de que perdemos de vista para qué sirve una definición. Las definiciones no son la verdad, tan sólo establecen las condiciones bajo las cuales han de tomarse acciones particulares (Farnham-Diggory, 1978).

Las acciones a las que se hace referencia pueden incluir investigaciones, evaluaciones diagnósticas, intervenciones educativas o todas. Pero cualquiera que sea la respuesta que se ponga en práctica, uno debe protegerse contra la propensión a considerar las definiciones como verdaderas o falsas; son únicamente invenciones humanas que tienen mayor o menor utilidad.

CARACTERÍSTICAS ESENCIALES DE LAS DISCAPACIDADES PARA EL APRENDIZAJE

Aunque la definición de las DA sigue siendo punto de discusión, se pueden citar varias características de estos trastornos que son menos polémicas. Como descubrirá el lector, las particularidades que se analizan en las siguientes páginas dictan, hasta cierto grado, la naturaleza de las prácticas de prueba en la evaluación de las DA. Existe un acuerdo general —con ocasionales votos en contra— sobre las siguientes características de estos padecimientos:

1. Una DA implica una discrepancia en el funcionamiento cognitivo. El niño (o adulto) con DA revela una debilidad relativa en un área en comparación con fortalezas en la mayor parte de las otras áreas. Según la definición federal que se sigue dentro de muchos sistemas escolares, la discrepancia existe entre la capacidad general (inteligencia) y el rendimiento específico. Antes ya se han descrito algunos de los escollos de esta definición y se prefiere el enfoque del NJCLD, en el que la discrepancia no se vincula de manera rígida con una diferencia entre las puntuaciones de CI y de una prueba de rendimiento.

2. En la mayor parte de las definiciones de DA se incluye una cláusula de exclusión. Si las dificultades académicas son principalmente el producto de otros padecimientos discapacitantes (retraso mental, trastorno emocional, deterioro visual o auditivo, desventajas culturales o sociales), entonces es común descartar el diagnóstico. Con frecuencia esta cláusula se interpreta de manera errónea. Una persona puede tener tanto una DA como otro tipo de trastorno (p. ej., retraso mental). El elemento importante es que el padecimiento coexistente no sea la causa principal de las DA.
3. Las DA son heterogéneas, es decir, son diferentes. La investigación acerca de la identificación de los subtipos todavía está en su infancia, pero la mayoría de los investigadores expresan optimismo en que puedan identificarse subgrupos significativos de personas con DA. Pendiente de mayor investigación y refinamiento, en la actualidad sólo se reconocen dos categorías amplias de estos trastornos (Forster, 1994):

- Dislexia o discapacidad para el aprendizaje verbal
- Discapacidad del hemisferio derecho o para el aprendizaje no verbal

En el cuadro 10–3 se resumen las características de estas dos categorías principales de DA. Estos patrones han surgido en muchos estudios sobre niños con DA. Por ejemplo, Blackely, Crinella, Fisher, Champaigne y Beck (1994) distinguieron tanto una DA verbal como una DA no verbal en un análisis complejo de las calificaciones de una prueba neuropsicológica con 177 niños de 9 a 14 años de edad. La muestra consistió en 129 niños con DA (incluyendo 37 con daño cerebral verificado) y 48 en quienes no había evidencia de DA o daño cerebral. Se

Cuadro 10–3. Características de dos categorías amplias de discapacidad para el aprendizaje

Dislexia o discapacidad para el aprendizaje verbal	Discapacidad del hemisferior derecho o del aprendizaje no verbal
Manifestación principal	
Dificultad inesperada en el aprendizaje de lectura u ortografía	Habilidades deficientes en matemáticas, escritura manuscrita o cognición social
Deficiencia fundamental	
Problemas en codificación fonológica (asociación de sonidos con combinaciones de letras)	Problemas en relaciones espaciales (percepción visoespacial de relaciones)
Correlatos fisiológicos	
Anomalías sutiles en el hemisferio cerebral izquierdo (revelados por rastreos cerebrales y estudios de EEG)	Probable origen en una disfunción del hemisferio cerebral derecho
Incidencia relativa	
Cerca de 90% de todos los casos de DA	Cerca de 10% de todos los casos de DA
Proporción de niños a niñas	
3:1 o 4:1	1:1

Fuente: Basado en Forster, A. (1994). Learning disabilities. En R. J. Sternberg (ed.), *Encyclopedia of human intelligence*. Nueva York: Macmillan.

identificaron seis patrones de desempeño neuropsicológico por medio de un complejo método estadístico de racimos:

1. **CI muy bajo:** niños con CI muy bajo, pero que en otros sentidos no tenían calificaciones discrepantes de prueba.
2. **CI bajo:** niños con CI bajo, pero que en otros sentidos no tenían calificaciones discrepantes de prueba.
3. **Torpes/letárgicos:** niños cuyas calificaciones de prueba indican torpeza o letargo.
4. **Disfunción del lenguaje:** niños con calificaciones relativamente bajas en variables del lenguaje.
5. **Disfunción espacial:** niños con buena función verbal pero con problemas en orientación espacial.
6. **Sin déficit:** niños que no tenían deficiencias que pudieran detectarse.

Los grupos 4 y 5 corresponden a las principales categorías de DA listadas anteriormente (DA verbal y no verbal), mientras los otros grupos representan a la normalidad (grupo 6), la baja capacidad intelectual (grupos 1 y 2), o variantes de posible DA (grupo 3). Por supuesto, algunos de los niños en los grupos 1 y 2 podrían satisfacer los criterios de DA si se les evaluara con pruebas adicionales.

4. Una DA es un fenómeno del desarrollo que por lo general es evidente en la primera infancia y que puede persistir hasta la edad adulta. Aunque los esfuerzos de regularización deben basarse en el optimismo —a fin de evitar las profecías autocumplidas— también es necesaria cierta dosis de realismo. Los estudios con niños que sufren graves DA sugieren que la mejoría notable en rendimiento académico es la excepción más que la regla, incluso cuando estos individuos reciben una intervención educativa intensiva. Por ejemplo, Frauenheim y Heckerl (1983) volvieron a evaluar a 11 adultos a quienes se había diagnosticado con DA en su infancia. Todos los participantes habían recibido ayuda especial en lectura: nueve se graduaron de la educación media superior y dos terminaron el primero de preparatoria. Los CI Totales se ubicaron por lo común en la parte baja del rango de 90, con el CI Verbal por debajo del promedio (media de 85) y el CI de Ejecución por arriba del promedio (media de 104). A pesar de la intervención de regularización, cuando se les volvió a examinar en la edad adulta **exactamente** con la misma prueba de rendimiento, la *Wide Range Achievement Test* (Prueba de Rendimiento de Amplio Rango), estos individuos habían mejorado poco con respecto a sus resultados en la escuela primaria. Estos datos se corroboraron en otros estudios de seguimiento (para una revisión, véase Kolb y Whishaw, 1990; cap. 29). Dichos resultados indican que los especialistas que trabajan con niños que tienen DA no deben concentrarse únicamente en las preocupaciones académicas. Los problemas sociales y emocionales —que pueden ser más susceptibles a la intervención— también demandan atención.
5. Con frecuencia, los individuos con discapacidad para el aprendizaje experimentan trastornos sociales y emocionales que son tan generalizados y trascendentales como las dificultades en rendimiento académico. Estos problemas pueden persistir hasta la adolescencia y edad adulta. De hecho, las secuelas socioemocionales a menudo se vuelven la principal afección exhibida, lo cual puede complicar el proceso de prueba y oscurecer el diagnóstico. Por ejemplo, en un estudio sobre evaluación de necesidades con 381 adultos con DA, Hoffman, Sheldon, Minskoff y colaboradores (1987) identificaron varias áreas importantes, no académicas, que merecían intervención por parte de los encargados de servicios. Estos adultos expresaban con gran frecuencia varios problemas sociales y emocionales: sentimientos de frustración (40%), hablar o actuar antes de pensar (33%), timidez (31%), falta de confianza en sí mismos (28%), control de emociones y carácter (28%) y salir con personas en una situación romántica (27%). También se expresaron muchos otros problemas, pero en menos de 25% de la muestra. Estos datos

indican que las evaluaciones de la discapacidad para el aprendizaje deberían incorporar medidas del funcionamiento social y emocional. Vaughn y Haager (1994) proporcionan una excelente revisión general sobre las habilidades sociales en personas con discapacidad para el aprendizaje.

CAUSAS Y CORRELATOS DE LAS DISCAPACIDADES PARA EL APRENDIZAJE

Aproximadamente 4% de todos los niños en edad escolar reciben un diagnóstico de DA, de modo que éste no es un problema poco frecuente (Chalfant, 1989). El tipo más común de DA es la dislexia, y los niños superan en número a las niñas en una proporción de cerca de 3:2 (Nass, 1992). En una minoría de casos, la etiología es clara y puede atribuirse a una causa específica como una lesión cerebral conocida. El lector recordará del capítulo 9 que el deterioro en el hemisferio izquierdo tiene especial probabilidad de resultar en dificultades verbales, mientras que el daño en el hemisferio derecho puede conducir a problemas con el pensamiento espacial y otras habilidades no verbales. Así, la lesión cerebral u otros problemas neurológicos pueden ser la causa inmediata de que un niño reciba un diagnóstico de DA.

No obstante, en la mayor parte de los casos, la etiología directa de estos problemas es poco clara. Se han propuesto varias posibilidades que pueden explicar sólo algunos de ellos. Por ejemplo, se han identificado procesos patológicos del desarrollo neurológico en algunas personas con dislexia grave (Culbertson y Edmonds, 1996). Los individuos con este trastorno parecen tener alteraciones en estructuras cerebrales como el plano temporal (la superficie plana en la parte superior de los lóbulos temporales), del cual se sabe que es importante para el procesamiento del lenguaje. Mientras que en los individuos normales el plano temporal es mucho más grande en el lóbulo temporal izquierdo que en el derecho, las personas con dislexia grave no muestran este patrón de asimetría (más bien una tendencia a la simetría). Lo que es más, los investigadores han identificado malformaciones corticales microscópicas llamadas polimicrogiria (numerosas circunvoluciones pequeñas) análogas a estas diferencias estructurales. Varios estudios *post mortem* de personas con dislexia grave han revelado estas desviaciones a nivel celular. La dislexia también parece mostrar un importante componente genético en el caso de algunas personas, de modo que necesita tomarse en serio la idea de dislexia familiar. Sin embargo, es necesario enfatizar que para la mayoría de los individuos, la etiología de la DA (ya sea dislexia u otras formas) aún es un misterio.

EVALUACIÓN DE LAS DISCAPACIDADES PARA EL APRENDIZAJE

Las discapacidades para el aprendizaje se manifiestan principalmente como problemas académicos; es decir, es común que un niño con DA no pueda dominar las habilidades importantes para el éxito escolar como la lectura, matemáticas o comunicación escrita. Debido a que los logros en la escuela están en el centro del problema, una valoración de la DA debe incluir medidas relacionadas con el rendimiento académico. Además, esta valoración —una pequeña parte de la evaluación de una DA— debe basarse en una prueba **individual** de rendimiento. Aunque una prueba grupal de rendimiento podría hacer surgir la sospecha de una DA, los profesionales deben depender de las pruebas individuales para la evaluación definitiva. A continuación se explican las razones para esto y después se revisan los instrumentos útiles para la evaluación del rendimiento.

Por lo común, las pruebas individuales de rendimiento se aplican en una situación frente a frente, donde el examinador se sienta de cara a la persona evaluada y le plantea preguntas y problemas. Por supuesto, cualquier prueba bien estandarizada del rendimiento producirá datos normativos acerca del funcionamiento de un escolar, pero la virtud especial de estas pruebas es que el examinador puede observar los

detalles clínicos del desempeño deficiente (o superior) y formar hipótesis acerca de las capacidades cognitivas de la persona examinada.

Considérese el problema de la ortografía deficiente, que se observa mucho en niños y adultos con DA verbal. Cualquier prueba buena de rendimiento en ortografía documentará la discapacidad; no obstante, se obtiene poco discernimiento a partir de las meras calificaciones. Lo que el examinador debe tratar de conocer es la naturaleza cualitativa del problema, no sólo sus dimensiones cuantitativas. Las pruebas individuales de rendimiento son inapreciables en ese sentido. Mediante la observación de los detalles del desempeño deficiente, un examinador hábil puede formular hipótesis acerca del origen de un problema de rendimiento. Por ejemplo, un niño cuya ortografía es fonéticamente correcta, al menos está **escuchando** las palabras de manera adecuada, mientras que uno con ortografía fonética incorrecta bien podría revelar un problema con el procesamiento auditivo de los sonidos del lenguaje.

PRUEBAS INDIVIDUALES DE RENDIMIENTO

Existe más de una docena de pruebas de rendimiento de aplicación individual, pero sólo pocas se utilizan de manera general en la evaluación de las DA. Las pruebas individuales de rendimiento más importantes incluyen *Diagnostic Achievement Batery –Second Edition*, DAB–2 (Batería de Diagnóstico del Rendimiento–2); *Kaufman Test of Educational Achievement*, K-TEA (Prueba Kaufman de Rendimiento Educativo); *Mini-Battery of Achievement*, MBA (Minibatería de Rendimiento); *Peabody Individual Achievement Test –Revised*, PIAT-R (Prueba Peabody de Ejecución Individual –Revisada); *Wechsler Individual Achievement Test*, WIAT (Prueba Wechsler de Rendimiento Individual) y Batería Psicoeducativa Woodcock-Johnson –Revisada (WJ-R), y la *Wide Range Achievement Test*, WRAT–III (Prueba de Rendimiento de Amplio Rango–III). Las características esenciales de estas pruebas se resumen en el cuadro 10–4. Debido a las limitaciones de espacio, se seleccionará una prueba, la K-TEA, para una presentación más detallada. Se sugiere a los lectores que deseen mayor información sobre estas pruebas que consulten a Sattler (1988, cap. 13) y la serie del *Mental Measurements Yearbook* (Conoley e Impara, 1995; Conoley y Kramer, 1989, 1992; Mitchell, 1985).

KAUFMAN TEST OF EDUCATIONAL ACHIEVEMENT

La K-TEA es una prueba sin límite de tiempo del rendimiento educativo para niños entre 6 y 18 años de edad. Existe una versión breve con tres subpruebas, pero cuando se trata de una evaluación diagnóstica de las DA se prefiere la Forma Amplia. Ésta incluye cinco subpruebas: Decodificación de Lectura, Comprensión de Lectura, Aplicaciones Matemáticas, Cálculos Matemáticos y Ortografía. El tiempo de la prueba es de aproximadamente una hora para los niños de mayor edad, y un poco menos para aquellos más pequeños.

En el cuadro10–5 se muestran ejemplos breves de reactivos similares a los de la K-TEA. Éstos se encontrarían en el extremo superior de las subpruebas y serían apropiados para estudiantes de educación media superior. La K-TEA utiliza reglas de inicio y terminación para cada subprueba, a fin de asegurar que los estudiantes sólo tengan que responder a reactivos de dificultad apropiada. La calificación es completamente objetiva: 1 para los reactivos correctos y 0 para los incorrectos. Las puntuaciones en bruto se convierten en puntuaciones estándar (media de 100, DE de 15) para cada subprueba, para la puntuación compuesta de lectura, puntuación compuesta de matemáticas y para la puntuación compuesta de la batería completa.

Además de la calificación formal, la K-TEA proporciona un método sistemático para evaluar la naturaleza cualitativa de los errores de subprueba. Por ejemplo, en la subprueba de Ortografía, los errores pueden clasificarse si implican prefijos, sufijos, dígrafos de vocal

Cuadro 10–4. Revisión de pruebas individuales de rendimiento de más uso

***Diagnostic Achievement Battery–2,* DAB-2 (Batería de diagnóstico de rendimiento)**

Newcomer (1990)

Adecuada para edades de 6 a 14 años, la DAB-2 consiste en 12 subpruebas que se emplean para calcular ocho puntuaciones compuestas mixtas. Las puntuaciones compuestas incluyen Audición, Habla, Lectura, Escritura, Matemáticas, Lenguaje hablado, Lenguaje escrito y Total de Aprovechamiento. Más amplia que la mayor parte de las pruebas de rendimiento, la DAB-2 requiere hasta dos horas para su aplicación. Las normas de la prueba se establecieron cuidadosamente con 2 600 niños a nivel nacional.

***Kaufman Test of Educational Achievement,* K-TEA (Prueba Kaufman de Rendimiento Educativo)**

Kaufman y Kaufman (1985)

La K-TEA es una prueba individual de rendimiento con normas adecuadas, una de cuyas características especiales es el detallado análisis de errores (véase texto). En la actualidad, las normas se extienden sólo hasta las edades de bachillerato. Una forma breve, independiente, que puede aplicarse en 30 minutos o menos, es útil para propósitos de detección.

***Mini-Battery of Achievement,* MBA (Minibatería de Rendimiento)**

Woodcock, McGrew y Werder (1994)

Evalúa cuatro amplias áreas de rendimiento —lectura, escritura, matemáticas y conocimientos objetivos— para personas de 4 a 90 años de edad o más. La batería completa puede aplicarse en 30 minutos. La MBA proporciona una cobertura más general de las habilidades básicas y aplicadas que ninguna otra batería breve. Por ejemplo, el componente de lectura evalúa identificación de letra-palabra, vocabulario y comprensión.

***Peabody Individual Achievement Test –Revised,* PIAT-R (Prueba Peabody de Ejecución Individual –Revisada)**

Markwardt (1989)

Para edades de 5 a 18 años, esta prueba de 60 minutos incluye subpruebas de información general, reconocimiento de lectura, comprensión de lectura, matemáticas y ortografía. En la actualidad se ofrece una nueva subprueba, expresión escrita, para detectar habilidades de lenguaje escrito. Su aplicación requiere de entrenamiento mínimo; la pueden aplicar maestros con entrenamiento apropiado.

***Wechsler Individual Achievement Test,* WIAT (Prueba Wechsler de Rendimiento Individual)**

Wechsler (1992)

La Prueba Wechsler de Rendimiento Individual consiste en ocho subpruebas: lectura básica, razonamiento matemático, ortografía, comprensión de lectura, operaciones numéricas, expresión oral, comprensión auditiva y expresión escrita. Adecuada para edades de 5 a 19 años, las normas de la WIAT se obtuvieron junto con las del WISC-III. La aplicación a personas mayores puede requerir hasta 75 minutos. Una versión más corta, el Detector WIAT (lectura básica, razonamiento matemático, ortografía) puede aplicarse en 10 a 15 minutos.

Woodcock-Johnson –Revisada, WJ-R

Woodcock y Johnson (1989)

La Woodcock-Johnson –Revisada cubre a individuos desde los dos años de edad hasta la edad adulta. La batería completa comprende tres áreas de funcionamiento: rendimiento, capacidad cognitiva e intereses. Las nueve pruebas estándar de rendimiento producen puntuaciones agrupadas en áreas denominadas Lectura Amplia, Matemáticas Amplias, Lenguaje Escrito Amplio, Conocimiento Amplio y Habilidades. Las pruebas de rendimiento son sumamente respetadas, pero algunos críticos recomiendan tener precaución en el uso de la batería cognitiva.

***Wide Range Achievement Test,* WRAT-III (Prueba de Rendimiento de Amplio Rango–III)**

Wilkinson (1993)

La WRAT-III tiene normas adecuadas para edades desde los 5 hasta los 75 años y se le utiliza de manera amplia como instrumento de detección. Las subpruebas incluyen lectura, ortografía y aritmética. La principal debilidad de la batería es la subprueba de lectura, la cual en realidad sólo es una medida del reconocimiento de palabras. Dicha subprueba consiste en pedirle a la persona examinada que pronuncie en voz alta cada palabra de una lista que va desde palabras sencillas hasta difíciles. Debido al limitado contenido de los reactivos y a las altas correlaciones entre subpruebas, la WRAT-III es inadecuada para la identificación de deficiencias en habilidades específicas.

Cuadro 10–5. Ejemplos de reactivos característicos de la K-TEA aplicables a niños mayores

Decodificación de lectura

El examinador señala cada palabra a la vez y dice: "¿qué palabra es ésta?"

dodecágono vociferante correlativo
indolencia perspicacia

Comprensión de lectura

El examinador dice: "Haz lo que esto dice".

Expresa una respuesta falsa para la pregunta: "¿cuántos ojos tiene un cíclope?"

Ortografía

El examinador explica las reglas para un texto tradicional de ortografía, concluyendo con: "Quiero que escribas la palabra en esta hoja".

"Bienquerida. Una persona amada es Bienquerida"

Cálculo matemático

El examinador dice: "Ahora quiero que resuelvas estos problemas".

$(X - 7)(X - 9) =$

$$\begin{array}{rr} 5\ \text{Kg} & 5\ \text{gr} \\ -2\ \text{Kg} & 14\ \text{gr} \\ \hline \end{array}$$

Aplicaciones matemáticas

El examinador dice: "Los Diablos de Durango jugaron 80 partidos el año pasado. Ganaron 16 juegos. ¿Qué porcentaje de juegos ganaron?"

(p. ej., ue en *blue**) y diptongos, agrupaciones de consonantes (como scr en inescrupuloso), patrones controlados por r (como er en inferior) y otros patrones.

Kaufman y Kaufman (1985) destacan que el análisis de los errores le proporciona al profesional del diagnóstico una fuente de información a partir de la cual pueden elaborarse los objetivos educativos. Por ejemplo, una debilidad en dígrafos y diptongos de vocal en la subprueba de Ortografía se traduce de manera directa en objetivos para el salón de clase: práctica de ortografía y lectura de estos elementos en aislamiento, progresión a ortografía y pronunciación de palabras que contengan dígrafos y diptongos y terminación con escritura y lectura de oraciones que contengan palabras con dígrafos y diptongos. El Manual de la K-TEA contiene muchos discernimientos clínicos útiles que tienen ramificaciones educativas.

Aunque las muestras normativas de la K-TEA y la descripción de sus características son un modelo de excelencia, las características técnicas de este instrumento varían en cuanto a suficiencia. Las confiabilidades por mitades para las cinco subpruebas van de .87 a .96, que son valores bastante aceptables para una prueba de rendimiento. Los datos de estabilidad son menos claros. Los datos para 172 estudiantes a los que se volvió a evaluar a los 35 días se desplomaron en el caso de los grados 1 a 6 y secundaria a bachillerato. Todas las correlaciones de subpruebas y compuestos superaron al .90, pero es probable que estos valores estén inflados debido a que mezclan la estabilidad del rendimiento con las correlaciones de grado contra rendimiento.

La validez de contenido de la K-TEA parece ser muy fuerte, pero este aspecto puede variar de un sistema escolar a otro. Después de todo, cada sistema escolar puede optar por enfatizar diferentes áreas de rendimiento. Salvia e Ysseldike (1991) advierten que los usuarios deben ser sensibles a la correspondencia entre el contenido de la K-TEA y el currículo de los estudiantes. Como ocurre con cualquier prueba de este tipo, el usuario debería verificar que el contenido de la K-TEA sea apropiado dentro del entorno académico. No obstante, Kaufman y Kaufman (1985) ofrecen suficiente evidencia de la validez de la prueba como para defender su idoneidad general.

BATERÍAS DE PRUEBA EN LA EVALUACIÓN DE LA DISCAPACIDAD PARA EL APRENDIZAJE

Aunque los expertos concuerdan en que la evaluación de una posible DA requiere de un enfoque multifacético, existe poco consenso en

* **Nota del Traductor.** En español se considera que las letras **ll** y **ch** son dígrafos. Sin embargo, no existen en realidad dígrafos de vocal como ocurre en otros idiomas (como en el francés **ou** —por ejemplo, "gourmet").

cuanto a cuáles son los mejores instrumentos y técnicas. Por supuesto, las herramientas esenciales en la evaluación de niños con esta discapacidad son las medidas confiables y válidas de inteligencia y rendimiento. Casi todas las baterías para DA incluyen instrumentos convencionales en ambas áreas; por ejemplo, la SB:CE, WPPSI-R o WISC-III para la evaluación intelectual y la PIAT-R, K-TEA, WJ-R o WIAT para la medición del rendimiento. Sin embargo, la elección de medidas auxiliares para evaluar habilidades del lenguaje, formas específicas de procesamiento de información o procesamiento visoperceptual diferirán de un profesional a otro. Lo que es más, muchos examinadores individualizarán la batería de prueba en función de las preguntas de canalización. En el ejemplo de caso 10–1 se proporciona una batería representativa de prueba para la evaluación de DA de Jaime, un niño de nueve años con antecedentes de fracaso en la escuela.

EVALUACIÓN DEL TRASTORNO POR DÉFICIT DE ATENCIÓN CON HIPERACTIVIDAD (TDAH)

Este trastorno es el término propuesto por la *American Psychiatric Association* para nombrar un síndrome conductual conocido anteriormente como trastorno por déficit de atención con hiperactividad, disfunción cerebral mínima e hiperquinesia (*American Psychiatric Association*, 1994). Es frecuente que los niños con este trastorno exhiban un rendimiento académico bajo. Por tanto, a menudo se les canaliza para una evaluación de DA. No obstante, aunque a menudo la DA y el TDAH coexisten y sus síntomas se superponen ligeramente, son síndromes distintos a un nivel conceptual. En nuestro caso, se revisarán ciertos instrumentos útiles diseñados para ayudar a diagnosticar el síndrome de TDAH.

El TDAH se presenta en tres variedades: predominancia de problemas de atención, predominancia de hiperactividad-impulsividad y tipo combinado. El primero se define por la presencia de seis o más síntomas de distracción, pero menos de seis síntomas de hiperactividad-impulsividad. El tipo hiperactivo-impulsivo se define por la presencia de seis o más síntomas de hiperactividad-impulsividad, pero menos de seis síntomas de distracción. El tipo combinado consiste en seis o más síntomas de ambos grupos. En todos los casos, los síntomas deben estar presentes al menos durante seis meses, y deben conducir a un deterioro del funcionamiento social, académico o laboral. Los síntomas diagnósticos del TDAH se resumen en el cuadro 10–6.

Cuadro 10–6. Síntomas diagnósticos del Trastorno por Déficit de Atención con Hiperactividad

Tipo con problemas de atención (seis o más síntomas)
- Falta de atención a los detalles
- Dificultad para mantener la atención
- No parece escuchar
- Fracaso en la consecución de una tarea
- Dificultad para organizar tareas
- Evita el esfuerzo mental sostenido
- Pierde las cosas
- Se distrae con facilidad
- Olvidadizo en actividades cotidianas

Tipo hiperactivo-impulsivo (seis o más síntomas)
- Inquietud, movimientos constantes o ambos
- Abandona su asiento en clase
- Corre o se sube en objetos de manera inapropiada
- Dificultad para jugar tranquilamente
- Parece vehemente, siempre en movimiento
- Habla mucho
- Responde de manera abrupta
- Dificultad para esperar su turno
- Interrumpe a los demás o se entromete con ellos

Tipo combinado
- Seis o más de los síntomas en cada una de las áreas anteriores

Fuente: Basado en la American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4a ed.). Washington, D. C.: Autor.

EJEMPLO DE CASO 10–1
UNA BATERÍA DE PRUEBA EN LA EVALUACIÓN DE LA DISCAPACIDAD PARA EL APRENDIZAJE

Jaime es un niño de nueve años de edad canalizado por su pediatra para valoración de una posible discapacidad para el aprendizaje. Sus padres y maestros caracterizaron a Jaime como un niño distraído, tranquilo y encantador, pero flojo en cuanto a su trabajo escolar, en particular en relación con las tareas escritas. Con frecuencia parecía que no prestaba atención a lo que el maestro decía. El propósito de la evaluación era determinar si en el caso de Jaime se podía diagnosticar una DA y también ofrecer recomendaciones para regularización.

El examinador utilizó instrumentos estándar como el WISC-III y la Batería Psicoeducativa Woodcock-Johnson –Revisada (WJ-R), pero también los complementó con medidas diseñadas para las preocupaciones específicas de la canalización (p. ej., problemas con distracción y comprensión auditiva). Lo que reveló la batería de pruebas fue una imagen de capacidad intelectual adecuada en términos generales (CI Total del WISC-III de 100), con relativa debilidad en áreas verbales (p. ej., Índice de Comprensión Verbal de 89) y capacidad relativa en áreas no verbales como organización perceptual, memoria visual e integración visomotora (p. ej., Índice de Organización Perceptual de 110). No obstante, a pesar de estas áreas de competencia, Jaime funcionaba a un nivel mucho menor en, cuando menos, un área académica (Muestras de escritura en el percentil 5). Estos datos corroboraron con toda seguridad las impresiones del maestro y de los padres. A nivel descriptivo, estos resultados sugirieron fuertemente un diagnóstico de discapacidad para el aprendizaje verbal, en especial, dado que podían descartarse otros factores contribuyentes (p. ej., trastorno emocional, instrucción deficiente).

No hay duda de que la fuente de sus dificultades tenía algo que ver con sus capacidades sumamente deficientes en cuanto a comprensión auditiva. Jaime tenía un desempeño en el porcentaje más bajo para niños de su edad en la *Test for Auditory Comprehension of Language –Revised*, TACL-R (Prueba de Comprensión Auditiva del Lenguaje) y en la *Wepman Auditory Discrimination Test* (Prueba Wepman de Discriminación Auditiva). En la TACL-R, el examinando escucha una palabra u oración y después señala al dibujo en blanco y negro (de un grupo de tres) que mejor representa el significado. En la Prueba Wepman de Discriminación Auditiva, el examinador lee un par de palabras similares y el niño indica si son iguales o diferentes. Las dificultades en ambas pruebas representaron una fuerte indicación de que Jaime exhibía un déficit fundamental en el procesamiento y comprensión de palabras habladas. Este déficit ayudó a explicar su problema con la expresión escrita (Rango percentil de 5 en Muestras de escritura). Sin duda, sus dificultades de atención también contribuían al cúmulo de problemas que experimentaba.

Una conclusión importante que surgió de los resultados de prueba fue que Jaime podría aprender mejor a partir de instrucciones escritas que habladas. Después de todo, tuvo puntuaciones exactas en el rango promedio para habilidades de lectura (Comprensión de textos de 100). Otra implicación de los datos era que aunque mostraba ciertas dificultades de atención, éstas no eran tan graves como para demandar una valoración posterior. El principal centro de atención de la intervención consistió en moderar su dificultad muy significativa en comprensión auditiva.

EJEMPLO DE CASO 10–1
UNA BATERÍA DE PRUEBA EN LA EVALUACIÓN DE LA DISCAPACIDAD PARA EL APRENDIZAJE (continuación)

Resumen de los resultados de prueba de Jaime

	Equivalente en Puntuación Escalar	Percentil 0	10	30	50	70	90	100
Habilidades verbales								
Índice de comprensión verbal	89							24
Prueba de hallazgo de palabras	88							22
Habilidades de organización perceptual								
Índice de organización perceptual del WISC-III	110							74
Prueba de inteligencia no verbal–3	109							72
Integración visomotora								
Test Guestáltico Visomotor de Bender	120							91
Prueba del desarrollo de la IVM (Integración visomotora)	107							68
Comprensión auditiva								
TACL-R (Comprensión auditiva)	72							3
Prueba Wepman de Discriminación Auditiva	65							1
Memoria								
Índice de memoria verbal de la WRAML	82							12
Índice de memoria visual de la WEAML	99							48
Habilidades de atención								
Índice de distractibilidad del WISC-III	84							15
Tarea de vigilancia (Sistema Gordon)	82							12
Tarea de distractibilidad (Sistema Gordon)	80							9
Rendimiento académico								
Comprensión de Textos de la WJ-R	100							50
Muestras de escritura de la WJ-R	75							5
Cálculo de la WJ-R	95							37
Problemas aplicados de la WJ-R (Matemáticas)	110							74

Por desgracia, estos criterios oficiales son vagos al grado que llevan al desaliento y relativamente comunes incluso en niños normales. No es sorprendente que los estimados de hiperactividad abarquen de 3 a 20% de la población en edad escolar (Cantwell, 1975). Para complicar todavía más las cosas, los expertos en este campo enfatizan otros elementos en la imagen general del TDAH. Por ejemplo, Barkley (1981) señala que los niños con TDAH muestran deficiencias principalmente en situaciones donde las instrucciones requieren respuesta demorada, sostenida o ambas, de acuerdo con sistemas de reglas bien definidos. También enfatiza que estos niños tienen un desempeño deficiente cuando se demoran los reforzamientos. Así, en el ambiente común de prueba donde se dan tareas interesantes con recompensas inmediatas, un niño con TDAH puede parecer bastante normal. El examinador que sospecha que un niño exhibe TDAH enfrenta un reto diagnóstico abrumador. Por fortuna, varios sistemas confiables y válidos de calificación pueden ayudar en la elaboración del diagnóstico.

Conners (1990, 1991) ha producido una familia de escalas de calificación útiles para identificar la hiperactividad y otros problemas conductuales en niños. Las *Conners' Teacher Rating Scales*, CTRS (Escalas Conners de clasificación del Maestro), se pueden obtener en dos formas, una que contiene 28 reactivos (CTRS-28) y otra con 39 (CTRS-39). El maestro califica al niño en cada uno de los reactivos en una escala de cuatro puntos. Las dos formas producen puntuaciones escalares diferentes y poseen propiedades psicométricas diversas. Las normas de la CTRS-28 se establecieron con 383 niños de 3 a 17 años de edad y produce tres escalas: Problema Conductual, Pasividad con Distracción e Hiperactividad. Las normas de la CTRS-39 se obtuvieron con 9 583 niños de 4 a los 12 años de edad y produce seis escalas: Hiperactividad, Indulgencia Emocional Extrema, Conducta Asocial, Problema Conductual, Ansiedad-Pasividad y Problema de Atención-Fantasías.

Las *Conners' Parent Rating Scales*, CPRS (Escalas Conners de Clasificación de Padres) también pueden obtenerse en dos formas, una que contiene 48 reactivos (CPRS-48) y otra con 93 (CPRS-93). Las normas de la CPRS-48 se aplicaron a niños de 6 a 14 años de edad y produce cinco escalas: Problemas Conductuales, Síntomas Psicosomáticos, Ansiedad, Problemas de Aprendizaje e Impulsividad-Hiperactividad. La CPRS-93, cuyas normas se aplicaron a niños entre 3 y 17 años de edad, incluye ocho escalas: Trastorno Conductual, Inquietud con Desorganización, Síntomas Psicosomáticos, Conducta Antisocial, Ansiedad Temerosa, Problema de Aprendizaje-Inmadurez, Obsesividad e Hiperactividad-Inmadurez.

En todas las escalas Conners, padres, maestros o representantes de los padres califican los síntomas en una escala de cuatro puntos (0 a 3). El formato de las diversas escalas Conners es de la siguiente naturaleza:

	De ninguna manera	**Sólo un poco**	**Bastante**	**Mucho**
Llora con facilidad				
Inquieto e intranquilo				
Actúa sin pensar				
Desobedece a los adultos				
Se mete en problemas				
Fantasea				

Trites, Blouin y LaPrade (1982) llevaron a cabo un análisis factorial de la CTRS-39 con una muestra aleatoria estratificada de gran tamaño con 9 583 escolares canadienses. Los resultados produjeron seis factores. El primero —y con mucho el más destacado— fue el de Hiperactividad, que presentó cargas en 17 de los 39 reactivos y explicó 36% de la varianza. El factor de Hiperactividad también tuvo excelente consistencia interna (coeficiente alfa de .94). En otros estudios sobre listas de verificación de la conducta infantil, la hiperactividad surge

como el factor principal y más fuerte cuando se realizan análisis factoriales de los reactivos de las escalas (Trites, 1979).

La evidencia de la validez de las escalas Conners es sustancial e incluye los siguientes elementos (Martens, 1992):

- Las calificaciones de escala muestran cambios apropiados cuando se trata a los niños hiperactivos con medicamentos que se sabe que mejoran la atención.
- Las escalas de calificación poseen correlaciones positivas fuertes con otras escalas de calificación, calificaciones de compañeros y observaciones independientes.
- Las puntuaciones de escala discriminan de manera apropiada entre grupos diagnósticos.

A pesar de la voluminosa base de investigación que sustenta estos instrumentos, los revisores expresan preocupación acerca de las muestras de estandarización. Por ejemplo, las normas de la CPRS-48 parecen basarse en una muestra de 529 niños casi exclusivamente de raza blanca del área de Pittsburgh, que pueden no ser representativos de la población general (Martens, 1992; Oehler-Stinnett, 1992).

Los editores de pruebas han publicado en años recientes una cantidad casi abrumadora de listas de verificación del TDAH y de otros problemas conductuales de la infancia. La mayor parte de estos instrumentos están diseñados para que los utilicen padres y maestros en el contexto de la evaluación escolar. Por ejemplo, Achenbach (1991, 1992) ha publicado versiones revisadas de su *Child Behavior Checklist*, CBCL (Lista de Verificación Conductual para Niños) para padres-informantes, un instrumento sumamente respetado que evalúa problemas sociales y competencia académica en una amplia variedad de dominios conductuales. En el cuadro 10–7 se resumen varios instrumentos nuevos, más restringidos.

EVALUACIÓN DE TRASTORNOS EMOCIONALES Y CONDUCTUALES

Aunque la mayoría de los niños se muestran despreocupados y disfrutan ir a la escuela, todos los maestros citan casos similares al siguiente:

> Pedro era un niño de nueve años de edad que cursaba el tercer grado y cuyo desempeño académico era irregular. Sus padres expresaban preocupación sobre su conducta ansiosa y retraída

Cuadro 10–7. Dominios evaluados con las escalas de calificación para los Trastornos por Déficit de Atención y Trastornos Relacionados

	Escalas			
Dominios	**ACTeRS**	**ADDES**	**ADHDT**	**CAAS**
Inatención	*	*	*	*
Conducta/Agresividad				*
Hiperactividad	*	*	*	*
Impulsividad		*	*	*
Oposicional	*			
Habilidades sociales	*			

ACTeRS = *ADD-H: Comprehensive Teacher's Rating Scale–2nd Edition* (Escala Amplia de Calificación del Maestro-2a edición); (Ullman, Sleator y Sprague, 1988).

ADDES = *Attention Deficit Disorders Evaluation Scales* (Escalas de Evaluación de los Trastornos por Déficit de Atención; McCarney, 1989).

ADHDT = *Attention-Deficit Hyperactivity Disorder Test* (Prueba del Trastorno por Déficit de Atención con Hiperactividad; Gilliam, 1994).

CAAS = *Children's Attention and Adjustment Survey* (Examen de Atención y Adaptación Infantil; Lambert, Hartsough y Sandoval, 1990).

en el hogar. Su maestra notaba ciertas conductas sociales "extrañas", como nunca ver a los otros niños a los ojos, gritar y hacer sonidos raros en el patio de recreo y parecer demasiado ansioso por complacer a otros durante los juegos. Pedro también parecía demasiado preocupado por conservar sus libros, papeles y lápices en un orden rígido sobre su pupitre. Pasaba muchos minutos diariamente ordenando y volviendo a ordenar estos materiales.

¿Qué le sucede a Pedro? ¿Cuál es la función del psicólogo escolar en cuanto a tratar los problemas aparentemente emocionales de este niño?

Además de evaluar la discapacidad para el aprendizaje, los psicólogos escolares también llevan a cabo valoraciones para determinar si los niños satisfacen los criterios de un trastorno emocional o conductual grave. Desde 1975, con la Ley Pública 94-142, el Congreso de EUA estipuló que los niños con graves trastornos emocionales o conductuales eran elegibles para obtener los servicios especiales subvencionados de manera indirecta por el gobierno federal. De aquí que, para propósitos de identificación, subsidios y tratamiento, los psicólogos escolares tengan que determinar si los niños designados tienen un trastorno emocional o conductual.

La ley federal clasifica a estos niños como "con Trastorno Emocional Grave" (TEG). El proceso por el cual se identifica la presencia de un TEG en un estudiante implica entrevistas con padres y maestros, valoración con escalas de calificación conductual y observación directa en el salón de clases. La meta consiste en identificar a los niños que exhiben conductas, sentimientos o patrones de interacción social inapropiados. En este caso se centra la atención en la función de las escalas de calificación conductual dentro de este proceso, debido a que proporcionan un enfoque relativamente objetivo para la evaluación.

En una escala de calificación conductual para niños, se pide a informantes como padres y maestros, que califiquen a un niño en relación con conductas relativamente aisladas, como **gusta de estar solo, se mete en peleas, es parlanchín, propenso a accidentes, se lleva con los demás** y así sucesivamente. Las calificaciones pueden ser dicotómicas (si-no), pero se presentan más en un continuo de 3 o 4 puntos (p. ej., nunca, en ocasiones, con frecuencia, siempre). Los reactivos se agrupan en escalas derivadas a través de análisis factorial que producen percentiles u otras calificaciones en referencia a muestras de estandarización de niños razonablemente normales (es decir, no canalizados). Se han desarrollado varias docenas de escalas de calificación conductual según esta estrategia. En el cuadro 10–8 se presentan algunos de los instrumentos de esta naturaleza que tienen más utilización. En nuestro caso se centra la atención en una herramienta que posee fundamentos empíricos extensos, la *Child Behavior Checklist*, CBCL (Lista de Verificación Conductual para Niños; Achenbach, 1991).

CHILD BEHAVIOR CHECKLIST

La CBCL es una de las escalas más cuidadosamente diseñadas y desarrollada de modo más minucioso en toda la psicología clínica. De hecho, el instrumento se presenta en formas diferentes, dependiendo de la edad del niño y de si la han de responder los padres o maestros. En este caso se restringe el análisis a la CBCL/4-18, adecuada para los informes de los padres acerca de capacidades y problemas en niños de 4 a 18 años de edad (Achenbach, 1991). El origen de esta herramienta multidimensional data de 1966, cuando Achenbach analizó más de 600 historiales clínicos de niños para identificar síntomas aislados relativamente fáciles de observar y que también pueden considerarse generales más que demasiado específicos. La investigación y examen posteriores resultaron en más de 100 reactivos que conforman la parte de problemas conductuales de la CBCL. Éstos se califican con 0 (no es cierto hasta donde usted sabe), 1 (un tanto o en ocasiones cierto) o 2 (muy cierto o a menudo cierto). Además, el instrumento incluye reactivos que detectan competencia social en tres categorías amplias: actividades, social y escuela.

Con base en numerosos estudios analítico factoriales, Achenbach descubrió que los padres

Cuadro 10–8. Listado breve de escalas representativas de calificación de la conducta infantil

Revised Behavior Problem Checklist
(Lista de Verificación de Problemas Conductuales –Revisada)

Quay y Peterson (1983)

Esta escala consiste de 89 reactivos, cada uno de los cuales se califica en una escala de gravedad de tres puntos; ésta produce seis dimensiones primarias; problemas conductuales, agresión socializada, problemas de atención, ansiedad-retraimiento, conducta motora y exceso motor.

Bahavior Assessment System for Children
(Sistema de Evaluación Conductual para Niños)

Reynolds y Kamphaus (1992)

Esta escala es un instrumento global que incluye versiones para padres, maestro y niños; la escala produce puntuaciones en dominios amplios de externalización e internalización, así como en áreas específicas de contenido que incluyen agresión, hiperactividad, problemas conductuales, problemas de atención, depresión, ansiedad, retraimiento, somatización y habilidades sociales.

Home Situations Questionnaire
(Cuestionario de situaciones hogareñas)

Barkley y Edelbrock (1987)

Esta escala consiste en 16 reactivos que tratan sobre situaciones hogareñas en las que pueden ocurrir conductas de falta de obediencia (p. ej., al niño se le pide que realice tareas); los padres califican cada reactivo en una escala de nueve puntos.

Social Skills Rating System
(Sistema de Calificación de Habilidades Sociales)

Gresham y Elliott (1990)

Disponible en formas para padres, maestro y autocalificación, éste es un cuestionario de 55 reactivos que proporciona información específica en tres dominios (habilidades sociales, conductas problema y competencia académica).

tienden a describir los problemas de los niños en ocho dimensiones: conducta agresiva, ansiedad/depresión, problemas de atención, conducta delincuente, problemas sociales, padecimientos somáticos, problemas de pensamiento y retraimiento. Estos patrones de escala se derivaron de los análisis factoriales de las calificaciones para 4 455 niños con canalización clínica. Los resultados se expresan como percentiles en comparación con una muestra normativa de 2 368 niños no canalizados. Además de las puntuaciones para las escalas individuales, la CBCL produce una puntuación de Internalización (los problemas se internalizan), una puntuación de Externalización (los problemas se externalizan) y una puntuación Total Problema (que refleja una desadaptación general).

Considérense los resultados de prueba de Pedro, el alumno problemático de tercer grado que se describió antes. Un aspecto valioso de la CBCL es que pueden compararse las calificaciones de ambos padres como una especie de verificación de confiabilidad. Los rangos percentiles para las calificaciones de su madre y padre se listan, respectivamente, a continuación, donde las puntuaciones más altas indican un problema mayor:

Conducta agresiva	52, 65
Ansiedad/Depresión	98, 91
Problemas de atención	98, 98
Conducta delincuente	34, 45
Problemas sociales	93, 91
Padecimientos somáticos	88, 89
Problemas de pensamiento	98, 77
Retraimiento	95, 95
Puntuación total problema	98, 98

En general, los resultados se encuentran dentro del rango clínico (la puntuación Total Problema se encuentra en el percentil 98). Las escalas individuales indican que se percibe a Pedro como sumamente retraído, ansioso/deprimido, con posibles problemas de pensamiento (p. ej., pensamientos extraños o peculiares) y con una dificultad evidente para prestar atención. La evaluación posterior realizada por el psicólogo escolar reveló una inteligencia promedio (CI

Total de 94) con una enorme discrepancia entre capacidad verbal (CIV de 110) y capacidad de desempeño (CIE de 79). En entrevista, Pedro reveló un tipo de pensamiento vago, desordenado, con notable distracción. En general, los resultados indicaron que calificaba para un diagnóstico de TEG y que requería de servicios psicológicos especiales en la escuela.

EVALUACIÓN DE TALENTOS ESPECIALES

Un ideal de las sociedades occidentales es que cada persona debería recibir educación adecuada para satisfacer sus potencialidades. Este ideal no sólo es consistente con el igualitarismo prevaleciente, sino que también demuestra una política astuta en la medida en que la educación adecuada para las personas muy inteligentes paga elevados dividendos para la sociedad en general. Se considera talentoso al individuo que desarrolla conceptos científicos originales y eficaces; descubre la cura para los males de la humanidad; produce grandes obras de arte e inventa productos, herramientas y máquinas útiles. La identificación temprana de los niños dotados es esencial si se han de nutrir sus talentos en beneficio de todos.*

Por lo general, la designación de una persona como **talentosa** significa que tiene una capacidad extraordinaria en cierta área (Horowitz, 1994). Renzulli (1986) señala que dentro de esta definición general, los expertos han seguido dos categorías amplias de talento especial, que podrían denominarse "talentos escolares" y "talentos creativos-productivos". La primera variedad la representan los estudiantes que sobresalen en las metas académicas tradicionales como la escritura, matemáticas o ciencias. Podría describirse a estos niños como personas que poseen talentos intelectuales. En cuanto a esta variedad de talento, Renzulli (1986) señala que:

> Existe en diversos grados; se le puede identificar a través de técnicas estandarizadas de evaluación y, por tanto, debemos hacer todo lo que esté en nuestro poder por realizar modificaciones apropiadas para los estudiantes que tienen la capacidad para cubrir el material curricular regular a tasas y niveles de entendimiento avanzados (p. 57).

Este tipo de talento especial puede evaluarse con facilidad a través de pruebas de CI o de otro tipo de pruebas de capacidad y rendimiento. La segunda categoría de talentos especiales, los creativos-productivos, es más difícil de valorar. La identificación del talento dentro de esta área depende de procedimientos más subjetivos —pocas pruebas son adecuadas para este propósito.

Quizás ésta es una de las razones por las que los sistemas escolares a menudo restringen el concepto de talentos especiales al área intelectual y dependen de pruebas estandarizadas para la identificación de niños elegibles. El punto en el que la capacidad intelectual es suficientemente elevada como para clasificar al niño como superior al nivel mental o talentoso es, por supuesto, una decisión un tanto arbitraria. Un enfoque típico consiste en reservar la etiqueta de talento especial (y proporcionar acceso a mejores oportunidades educativas) para los estudiantes que obtienen puntuaciones en 1% superior dentro de pruebas estándar de inteligencia, como las escalas Stanford-Binet o Wechsler. Esto se traduce en un CI de aproximadamente 135 y superior. En resumen, un enfoque para la identificación de niños talentosos implica la nominación de un maestro o padre, la aplicación de una prueba individual de CI apropiada y después la selección de los niños para experiencias educativas enriquecidas, con base en una puntuación suficientemente elevada de una prueba.

Una extensión del enfoque basado en una prueba para la identificación de los talentos es-

* Debido a que éste es un libro sobre pruebas psicológicas, no se puede analizar las tendencias en cuanto al financiamiento de la educación para los individuos talentosos. No obstante, es desalentador observar que en el decenio de 1990, menos de **una décima parte de 1%** de todos los recursos federales gastados en la educación primaria y secundaria se dirigieron a programas para personas talentosas (Irwin, 1992).

colares o intelectuales es el empleo de un sistema de calificación cuantitativa que incorpore datos de prueba, calificaciones escolares y recomendaciones del maestro (cuadro 10–9). Este método proporciona una plataforma más estable para la identificación del talento intelectual.

CONCEPTO CREATIVO-PRODUCTIVO DE LOS TALENTOS ESPECIALES

Con base en la sospecha de que las pruebas tradicionales dependen demasiado del conocimiento, experiencia y contenido específicos, los educadores han propuesto un enfoque alternativo para la definición e identificación de niños talentosos. Los talentos creativos-productivos se refieren a aquellos que poseen niños (y adultos) que sobresalen en el desarrollo de productos y materiales originales (Sternberg y Zhang, 1995). Encontrar a estos niños obliga a "buscar por debajo de 3 a 5% superior en la curva normal de puntuaciones de CI" (Renzulli, 1986, p. 58). De hecho, pocas pruebas son útiles para localizar estos talentos. La identificación de estos niños depende en gran medida del juicio subjetivo de autoridades (incluyendo maestros y psicólogos) que deben aplicar definiciones conceptuales de los talentos especiales para las circunstancias específicas de los niños individuales.

La perspectiva sobre talentos especiales que ofrece Renzulli (1978, 1986) sirve para ilustrar este enfoque alternativo. Este autor le resta importancia a la función del CI, y en vez de ello define los talentos especiales como la confluencia de **tres** elementos:

1. Inteligencia superior al promedio.
2. Evidencia de creatividad.
3. Evidencia de compromiso con la tarea.

El primer elemento especifica una capacidad superior al promedio, pero no necesariamente sobresaliente o dotada. Según este punto de vista, un niño talentoso es aquel que posee capacidad

Cuadro 10–9. Pautas de muestra para la identificación de talentos intelectuales especiales

Categoría	Rango de puntuaciones	Valor Ponderado*
Datos de prueba de inteligencia	percentil 95–99	10
	92–94	8
	89–91	7
	86–88	6
	80–85	5
Datos de prueba de rendimiento	percentil 95–99	8
	92–94	7
	89–91	6
	86–88	5
	80–85	4
Calificaciones escolares	A o 96–99 o Superior	8
	B o 91–95 o Muy bien	7
	C o 80–90 o Bien	6
Calificaciones del maestro	El más prometedor en clase	5
	Estudiante excelente	4
	Superior al promedio	3
	Promedio	2

* Puntuación máxima posible = 31 y se necesitan 28 puntos para identificar a un estudiante como talentoso.
Fuente: Basado en Gallagher, J. J. y Courtright, R. D. (1986). The educational definition of giftedness and its policy implications. En R. J. Sternberg y J. E. Davidson (eds.), *Conceptions of giftedness*. Cambridge; Cambridge University Press.

general superior al promedio (digamos un CI de 115 o mayor) o un talento reconocido en un área específica, como química, ballet, matemáticas, composición musical, esculpido y fotografía. Mientras la capacidad general y muchos talentos específicos pueden medirse con pruebas estandarizadas, algunas áreas como las artísticas deben valorarse a través de técnicas de evaluación basadas en el desempeño.

El segundo elemento esencial es que los niños talentosos revelan brotes repentinos de creatividad en sus actividades, pero, ¿qué es la creatividad y cómo se le puede evaluar como un elemento de los talentos creativos-productivos? La evaluación de la creatividad ha fascinado a los psicólogos y educadores durante decenios —y también ha resultado ser un problema enigmático. Aunque en general los investigadores reconocen que la **creatividad** es un tanto diferente de la inteligencia, más allá de este punto fundamental existe poco acuerdo en cuanto a la naturaleza o evaluación de la misma (Wallach, 1985).

A lo largo de los años, la creatividad se ha definido como un proceso, una característica personal y un producto conductual (Amabile, 1983). Un ejemplo de la perspectiva de proceso es la idea de que las personas creativas sobresalen en un proceso cognitivo específico llamado **producción divergente:**

> La producción divergente se define como del tipo que se dispara en direcciones diferentes. Hace posibles los cambios de dirección en la solución de problemas y también conduce a una diversidad de respuestas, donde más de una de ellas puede ser aceptable (Guilford, 1959).

Un ejemplo de una medida de producción divergente es la *Consequences Test* (Prueba de Consecuencias; Guilford y Hoepfner, 1971). Algunas de estas medidas incluyen preguntas como: "¿cuáles serían las consecuencias si las nubes tuvieran cordeles colgando de ellas?"; "¿cuáles serían las consecuencias si las bombillas eléctricas costaran $100 cada una?"; "¿cuáles serían las consecuencias si los océanos se elevaran tres metros?". El número de respuestas dadas se considera como un índice de producción divergente que, a su vez, se considera como evidencia de creatividad. Estas pruebas han disfrutado de popularidad periódica, pero su valor como medidas de creatividad aún es cuestionable.

Otro punto de vista acerca de la creatividad es que los rasgos personales representan su probable presencia. Según esta perspectiva, la creatividad proviene del temperamento, motivación y carácter del individuo; esto sugeriría que existe una personalidad creativa. Harrington (1975) ha capturado una representación no del todo halagadora de la persona creativa en su escala de *Composite Creative Personality* (Personalidad Creativa Mixta). Esta prueba es una lista de verificación de adjetivos: las personas creativas se distinguen al calificar en sí mismas la presencia de ciertos rasgos que incluyen argumentador, asertivo, apresurado, perspicaz, rebelde, espontáneo y versátil, entre otros. Esta línea de investigación indica que las personas creativas se distinguen por intereses, actitudes y motivaciones y no sólo por la capacidad intelectual. No obstante, el vínculo es indirecto e imperfecto. En consecuencia, las medidas de personalidad en pocas ocasiones ayudan a la identificación de los estudiantes talentosos.

El último enfoque para la creatividad utiliza el producto como signo distintivo de esta capacidad. Según este punto de vista, las personas creativas producen cosas (ideas, inventos, escritos, producción artística, etc.) que satisfacen ciertos criterios. Por ejemplo, Jackson y Messick (1968) aplican cuatro criterios a un producto creativo.

- **Novedad:** los productos creativos son nuevos o al menos representan una aplicación nueva a lo que ya es familiar.
- **Idoneidad:** el producto debe ser apropiado para el contexto y no solamente novedoso.
- **Trascendencia de las restricciones:** un producto trasciende cuando va más allá de lo tradicional.
- **Fusión de significado**: al principio, el valor de los productos creativos puede no ser aparente; es posible que la importancia completa se aprecie sólo con el tiempo.

Estos criterios pueden utilizarse para identificar a niños que parecen demostrar creatividad, uno de los elementos de los talentos creativos-productivos; se requiere del juicio subjetivo para identificarlos —no existen pruebas o escalas de calificación disponibles para este propósito.

El tercer elemento esencial en los talentos creativos-productivos es la evidencia del compromiso con la tarea. En la dedicación y pasión que ponen en sus fines, los niños talentosos sorprenden a padres y maestros por igual. Dichos niños pasan voluntariamente tanto tiempo en la búsqueda de sus talentos como sus compañeros lo ocupan en ver televisión (Renzulli, 1986). En los estudios clínicos sobre niños talentosos, las cualidades de persistencia, resistencia, participación, perseverancia, trabajo intenso, dedicación y confianza en sí mismos, se mencionan una y otra vez (Terman y Oden, 1959). La evaluación del compromiso con la tarea requiere de un enfoque subjetivo, aunque una especificación operacional podría ser que el candidato elegible debe haber mostrado una pasión por su área de talento durante un periodo específico (digamos que, cuando menos, durante un año o más).

Recientemente, varios expertos han expresado preocupación porque el concepto de talento especial se haya definido desde una perspectiva cultural eurocentrista, de tal modo que se pasa por alto a muchos estudiantes de minorías sin que se les reconozca (Hamilton, 1993; Maker, 1996). La dependencia extrema en las puntuaciones de prueba se considera particularmente problemática en la medida en que las pruebas existentes son medios cuestionables de predicción del éxito en ambientes no académicos —en especial para grupos de minorías étnicas, culturales y lingüísticas. Como un remedio, Maker (1996) recomienda que los educadores utilicen prácticas de identificación que se orienten al proceso y se basen en el desempeño (en contraposición a las puntuaciones de prueba). Esta autora desarrolló un enfoque auténtico para la identificación de niños talentosos de minorías que se denomina *DISCOVER* , en el que grupos de alumnos de primaria resuelven problemas con cubos, rompecabezas de figuras geométricas, libros de acertijos y juguetes:

> El propósito es crear una situación de solución de problemas similar a la que podría ocurrir en un salón de clases apropiado para estudiantes talentosos de diversos antecedentes culturales, étnicos, económicos y lingüísticos.

Los niños pasan por tres o más actividades, mientras los observadores documentan los procesos y habilidades de solución de problemas. Después, los observadores llenan una lista de verificación con 82 reactivos que tiene el propósito de identificar a los estudiantes talentosos. Debido a que los niños utilizan materiales concretos, pueden expresar lo que Gardner (1992) denomina conocimiento de "primer orden", el cual implica la creación y comprensión de historias, música, dibujos, construcciones y explicaciones. Este tipo de conocimiento depende menos del aprendizaje y rendimiento académicos en el sistema simbólico que se enseña en la escuela y, por tanto, se le considera como un índice más preciso de las capacidades de niños que provienen de antecedentes diversos. El enfoque *DISCOVER* recuerda que el talento especial puede verse desde muchas perspectivas —no necesariamente es una cosa.

RESUMEN

1. La detección de la preparación para la escuela es una importante función de la evaluación dentro de los sistemas escolares. Una prueba útil de detección debería producir tasas de falsos negativos menores a 20%, y de falsos positivos de menos de 10%. Por desgracia, los instrumentos de detección existentes pocas veces satisfacen estos criterios.
2. Instrumentos útiles para la detección en educación preescolar incluyen los DIAL-R, que evalúan habilidades motoras, conceptos

cognitivos y habilidades del lenguaje; la Denver II, que evalúa el desarrollo en cuatro áreas-personal-social, motora fina-adaptativa, lenguaje y motora gruesa; y la HOME.

3. La HOME es un índice del ambiente del niño que se basa en la observación dentro del hogar y la entrevista con la principal persona encargada del cuidado del niño. El inventario mide la calidad y cantidad de estimulación y apoyo para el desarrollo cognitivo, social y emocional que el niño tiene a su disposición en el hogar.
4. A corto plazo, es probable que no exista un mejor índice del riesgo que corre un niño de fracasar en la escuela que una prueba individual de inteligencia como la o las DAS.
5. Con base en la Ley Pública 101-476, una extensión de la Ley Pública 94-142, la definición federal de DA (definidas como una discrepancia significativa entre inteligencia y aprovechamiento) ha perdido el favor de los expertos en el campo de la DA. Una definición más novedosa hace referencia a una debilidad intraindividual en una o más áreas centrales de funcionamiento académico (capacidades de audición, habla, lectura, escritura, razonamiento o matemáticas) como la característica esencial de la DA.
6. Existe un acuerdo general —con ocasionales votos en contra— acerca de las siguientes características de las discapacidades para el aprendizaje, éstas implican una discrepancia intraindividual en cuanto a funcionamiento cognitivo; una exclusión de otros padecimiento discapacitantes como la causa principal; heterogeneidad, es decir, la existencia de muchos subtipos diferentes; una continuidad en el desarrollo desde la infancia hasta la edad adulta; y una alta incidencia de consecuencias sociales y emocionales.
7. Una de las pruebas individuales representativas de rendimiento que se utiliza en la evaluación de la DA es la K-TEA, una prueba sin límite de tiempo para niños de 6 a 18 años de edad, que consiste en cinco subpruebas: Decodificación de Lectura, Comprensión de Lectura, Aplicaciones Matemáticas, Cálculos Matemáticos y Ortografía.
8. Las puntuaciones de subprueba de la K-TEA pueden convertirse a equivalentes de grado y puntuaciones estándar con media de 100 y DE de 15. También es posible realizar un análisis cualitativo de los errores en las subpruebas, lo cual mejora la utilidad educativa de la K-TEA. La estandarización, confiabilidad y validez de contenido parecen ser buenas, aunque sería deseable realizar estudios *test-retest* adicionales.
9. Debido a que el trastorno por déficit de atención con hiperactividad (TDAH) a menudo coexiste con la DA, los profesionales requieren de herramientas y conceptos para la evaluación de este trastorno. Los criterios del DSM-IV para el TDAH destacan inquietud, distracción, impulsividad, deficiencias de atención, habilidades sociales deficientes y falta de consideración de las consecuencias. Otros enfatizan como síntomas diagnósticos las deficiencias en respuesta demorada, sostenida o ambas.
10. Conners ha desarrollado una familia de escalas de calificación para la identificación de hiperactividad y de otros problemas conductuales en niños. Maestros, padres y encargados de atender a los niños califican los síntomas en una escala de cuatro puntos. Estos instrumentos y otras escalas de calificación como la *Child Behavior Checklist* de Achenbach son complementos útiles en la evaluación de la conducta problema en niños.
11. Un niño con trastorno emocional grave (TEG) exhibe conductas, sentimientos o patrones de interacción social inadecuados. Las escalas objetivas como la CBCL son útiles para la evaluación del TEG. Uno o ambos padres califican los reactivos en la CBCL con 0 (no cierto), 1 (un tanto cierto o en ocasiones cierto) o 2 (muy cierto o a menudo cierto).

12. Otra función del psicólogo escolar es evaluar los talentos especiales, que se refieren a la capacidad extraordinaria en alguna área. Es frecuente que se identifique a los niños elegibles a través de un alto CI en una prueba individual de inteligencia, pero pueden utilizarse otros enfoques. Algunos expertos hacen referencia a la inteligencia superior al promedio, creatividad y compromiso con la tarea como ingredientes importantes de los talentos especiales. Este enfoque se basa en gran medida en el juicio subjetivo de autoridades (administradores, psicólogos) para la identificación de los niños talentosos.

TÉRMINOS Y CONCEPTOS CLAVE

Detección p. 420
Sensibilidad p. 422
Especificidad p. 422
Discapacidad para el aprendizaje p. 428
Trastorno por déficit de atención con hiperactividad p. 438
Talentos especiales p. 445
Creatividad p. 447
Producción divergente p. 477

TEMA 10B
Aplicaciones forenses de la evaluación

Normas para el testimonio experto
Valoración de presunta simulación
Evaluación del estado mental para la defensa de insania
Competencia para estar sujeto a juicio
Predicción de la violencia y evaluación del riesgo
Valoración de custodia de los niños en el divorcio
Lesiones personales y testimonio relacionado
Interpretación de prueba de polígrafo
Controversia sobre el psicólogo como testigo experto
Resumen
Términos y conceptos clave

La psicología y el sistema legal han tenido una alianza larga e intranquila, caracterizada por la desconfianza de ambas partes. Dentro del sistema legal, abogados y jueces conservan una antipatía hacia el testimonio de los psicólogos, debido a una preocupación de que sus opiniones se basen en "ciencia falsa" (o quizás en ninguna ciencia en absoluto) y también debido a una creencia (no del todo infundada) de que algunos testigos expertos profesarán casi cualquier punto de vista que sirva a los intereses del acusado. Dentro de la profesión de la salud mental, los psicólogos consideran que el aspecto de antagonismo del testimonio en un tribunal —basado en la expectativa de opiniones de si-no que se expresan como certezas virtuales— es un campo imposible en el cual se busca la verdad sobre la conducta humana. Como descubrirá el lector, esta tensión esencial entre la ley y la psicología es un telón de fondo constante que moldea y da forma a la naturaleza de la práctica psicológica en los tribunales.

Para bien o para mal, los psicólogos sí rinden testimonio en casos judiciales, y el centro de atención de su testimonio a menudo tiene que ver con la interpretación de pruebas psicológicas y entrevistas de evaluación. ¿Cuándo

son admisibles en una corte los resultados de prueba y las opiniones psicológicas basadas en ellas? ¿Qué criterio siguen los jueces en la determinación de si el testimonio psicológico debe admitirse? Los psicólogos que se presentan a sí mismos como expertos y utilizan las pruebas para justificar sus opiniones, deben tener una sólida fundamentación en cuanto a las cuestiones legales que se relacionan con la evaluación. En este tema se examinará la relevancia de las normas legales para el testimonio que se basa en pruebas y valoraciones psicológicas. También se examinarán algunos instrumentos especializados útiles en la evaluación forense.

La función del psicólogo examinador puede confluir con el sistema legal en una multitud de formas. Pueden solicitarse los servicios del profesional para lo siguiente:

- Valoración de presunta simulación
- Evaluación del estado mental para la defensa de insania
- Determinación de la competencia para someterse a juicio
- Predicción de violencia y evaluación del riesgo
- Valoración de custodia de los niños en el divorcio
- Evaluación de lesiones personales
- Interpretación de prueba de polígrafo

Éstas son las principales aplicaciones de la práctica forense que se analizarán aquí. Una variedad de aplicaciones adicionales se analizan en Melton, Petrila, Poythress y Slobogin (1998).

Además de satisfacer las pautas generales de práctica ética que se requieren de cualquier clínico, los profesionales que ofrecen testimonio experto basado en pruebas psicológicas se enfrentarán con normas adicionales para la práctica únicas del sistema de jurisprudencia de EUA. Aquí se resumen las principales preocupaciones concernientes a las pruebas psicológicas y el testimonio en tribunales. El lector puede encontrar análisis más amplios sobre este tema en Melton *et al.* (1998) y Wrightsman, Nietzel y Fortune (1994).

Cada uno de los temas antes listados hace surgir preguntas únicas acerca de la función del psicólogo en un tribunal. No obstante, un asunto es común a todas las formas de testimonio en una corte: ¿Cuándo puede considerarse como testigo experto a un psicólogo? Se analiza esta cuestión general antes de regresar a las aplicaciones específicas de la valoración psicológica que confluyen con el sistema legal estadounidense.

NORMAS PARA EL TESTIMONIO EXPERTO

De la misma manera en que los psicólogos se ocupan de las cuestiones de normas y competencia, también lo hacen los abogados y jueces. La jurisprudencia estadounidense ha desarrollado diversas pautas para el testimonio en tribunales, incluyendo varios principios generales relacionados con el testimonio de un **testigo experto**. Estas normas se encuentran en *Federal Rules of Evidence* (1975) y se han confirmado a través de varios dictámenes judiciales. Se pueden resumir los principios del testimonio de expertos como sigue:

- El testigo debe ser un experto calificado. No todos los psicólogos a los que se pida atestiguar se les permitirá hacerlo; con base en un resumen de su educación, entrenamiento y experiencia, el juez decide si el testimonio de ese testigo es admisible.
- El testimonio debe referirse a un tema apropiado. En particular, el experto debe presentar información que se encuentre más allá del conocimiento y experiencia del jurado promedio.
- El valor de la evidencia en cuanto a determinar la culpa o inocencia debe superar su efecto perjudicial. Por ejemplo, si el testimonio del experto puede provocar confusiones en cuanto al tema a discutir o causar que los miembros del jurado prejuzguen la situación, en general no es admisible.
- El testimonio de un experto debe ir de acuerdo con una teoría explicativa aceptada en general. En la mayor parte de los tribunales,

las pautas sobre esta cuestión las proporciona el caso de **Frye contra EUA**, un caso judicial ocurrido en 1923 relacionado con aceptabilidad del testimonio de expertos.

En este caso, los abogados de un individuo acusado de homicidio intentaron introducir los resultados de una prueba de engaño a través de la medición de la presión sanguínea sistólica. El abogado ofreció la presencia de un experto para atestiguar en cuanto a los resultados de dicha prueba. Se afirmó que la activación emocionalmente inducida del sistema nervioso simpático produce aumento gradual de la presión sanguínea sistólica si la persona examinada intenta engañar al examinador. En otras palabras, el testigo experto afirmó que en el curso de un interrogatorio acerca de un crimen, el patrón de cambio en la presión sistólica podía utilizarse como una forma de prueba de detección de mentiras. La defensa deseaba que el experto atestiguara en apoyo a la inocencia del cliente. La parte acusadora estuvo en contra de ello y el Tribunal de Apelaciones del Distrito de Columbia avaló la objeción, declarando que:

> ... aunque los tribunales están más que dispuestos a admitir el testimonio deducido de un principio o descubrimiento científico bien reconocido, aquello en lo que se basa la deducción debe haberse establecido a un nivel suficiente como para tener la aceptación general en el campo particular al que pertenezca (citado en Blau, 1984).

El tribunal concluyó que la prueba de engaño por medio de la presión sanguínea sistólica no había obtenido aceptación entre las autoridades fisiológicas y psicológicas y, por tanto, se negó a admitir el testimonio del experto.

Según estas pautas, una prueba, inventario o técnica de evaluación debe haber estado disponible durante un periodo lo suficientemente largo como para tener antecedentes de aceptación general. Por tal razón, el testigo experto prudente elegirá instrumentos bien establecidos y ampliamente investigados como base para su testimonio, en vez de depender de pruebas de desarrollo reciente que podrían no soportar el contrainterrogatorio bajo las limitaciones establecidas por el caso **Frye contra EUA**.

En 1993, las normas para el testimonio experto se refinaron aún más a través de la decisión de la Suprema Corte en el caso de **Daubert contra Merrell Dow Pharmaceuticals**. La opinión por escrito de este tribunal añadió pautas amplias acerca de los factores a considerar en la ponderación del testimonio científico en juicios. En particular, la corte dictaminó que una opinión experta debe basarse en "una inferencia o afirmación ... derivada del método científico". Las ramificaciones de este dictamen para el testimonio experto de los psicólogos son poco claras; no obstante, algunos tribunales han utilizado el dictamen de **Daubert** como base para impugnar el testimonio de profesionales de la salud mental, incluyendo a psicólogos. Por ejemplo, un tribunal determinó que las valoraciones psicológicas de niños que han sufrido abuso sexual eran inadmisibles como evidencia. Parece probable que los tribunales demandarán cada vez mayores pruebas de que el testimonio de los psicólogos tiene una base científica estricta (Melton y cols., 1998).

VALORACIÓN DE PRESUNTA SIMULACIÓN

En la mayor parte de los ambientes, los psicólogos pueden suponer de manera segura que los clientes son razonablemente honestos acerca de su estado mental y emocional. Los clientes quieren contar sus historias y desean transmitirlas de manera correcta. En el peor de los casos, pueden exagerar un tanto los síntomas para persuadir al clínico de que la ayuda es merecida y necesaria. Sin embargo, el engaño y la manipulación abiertos son poco comunes —por la sencilla razón de que no es común que los clientes tengan incentivo para estas estrategias.

No obstante, las reglas de participación clínica se voltean de cabeza en los ambientes forenses. El cliente forense típico tiene mucho que ganar de una formulación del caso que enfatice la enfermedad y discapacidad. De hecho, el

contexto de la evaluación casi garantiza que los clientes tendrán el propósito de parecer "locos" o discapacitados, ya sea por exageración o (en ocasiones menos frecuentes) como estratagema. En la mente del cliente forense, la fabricación de síntomas puede servir como excusa para la conducta inaceptable (p. ej., para favorecer una defensa de insania), para influir sobre las recomendaciones de condena (p. ej., contra la pena capital) o para obtener indemnizaciones (p. ej., certificación de discapacidad). Las maquinaciones de estos clientes influyen claramente en la validez de las evaluaciones forenses. En el fondo de toda evaluación de este tipo ronda la perturbadora pregunta: ¿el cliente fue razonablemente honesto y directo?

El examinador forense debe hacer un juicio sobre la honradez de la imagen que presenta el cliente durante la valoración. Y, sin embargo, aunque el sentido común dicta que el examinador debe esperar cierto grado de engaño, para llegar a la conclusión de que el cliente ha simulado de manera consciente, se necesita tener precaución:

> Dados el potencial importante de engaño y las implicaciones para la validez de sus datos, los profesionales de la salud mental deben desarrollar un umbral bajo de sospecha de las respuestas engañosas. Al mismo tiempo, debido a que la etiqueta de "simulador" puede tener un peso considerable en quienes toman las decisiones legales y tiene el potencial de empañar todos los aspectos de la posición legal de la persona, las conclusiones acerca de que una persona está fingiendo no deben alcanzarse de manera apresurada (Melton y cols., 1998).

El método más común y venerable para identificar a clientes deshonestos es la entrevista clínica. No obstante, debe preferirse un enfoque más objetivo. La evaluación de la posible **simulación** a través de la entrevista depende del juicio del clínico (p. ej., "el cliente es inconsistente en su presentación de los síntomas y parece deseoso de estar enfermo, de modo que concluyo que está simulando"), el cual puede resultar erróneo. En contraste, un enfoque objetivo proporciona datos normativos, tasas de acierto y similares para la valoración. Esto no sólo podría mejorar la precisión de la evaluación, además, un enfoque más estandarizado debería encontrar mayor aceptabilidad en muchos sistemas judiciales.

Por desgracia, existen relativamente pocos enfoques objetivos para la evaluación de la simulación en clientes forenses. Un instrumento prometedor es la *Structured Interview of Reported Syntoms*, SIRS (Entrevista Estructurada de Síntomas Reportados), un inventario de entrevista con 172 reactivos diseñado de manera expresa para la valoración de la simulación (Rogers, Bagby y Dickens, 1992). El enfoque que se incorpora en la SIRS se basó en las estrategias identificadas en la literatura clínica como potencialmente útiles para detectar la simulación. Con el empleo de un método de entrevista estructurada, ésta se evalúa en ocho escalas principales:

- Síntomas raros (informe excesivo de síntomas poco frecuentes).
- Combinaciones de síntomas (síntomas psiquiátricos reales que raramente ocurren juntos).
- Síntomas improbables o absurdos (los síntomas revelan una cualidad fantástica).
- Síntomas flagrantes (alusión excesiva a signos obvios de trastorno mental).
- Síntomas sutiles (alusión excesiva a problemas cotidianos).
- Gravedad de los síntomas (los síntomas se expresan con una gravedad extrema, insoportable).
- Selectividad de los síntomas (alusión indiscriminada a problemas psiquiátricos).
- Síntomas informados contra observados (comparación de los síntomas observados e informados).

Además de las ocho escalas primarias, se utilizan cinco complementarias para interpretar estilos de respuesta. De las 172 preguntas, 32 son interrogaciones repetidas para detectar inconsistencia en las respuestas. Ejemplos de estos tipos de preguntas de entrevista estructurada incluyen: "¿en alguna ocasión ha sentido como si las tapaduras de sus dientes pudieran captar mensajes de la radio?" (síntomas raros); "¿tie-

ne fuertes dolores de cabeza al mismo tiempo que teme a los microbios?" (combinaciones de síntomas); "¿el mobiliario del sitio donde vive parece hacerse más grande o más pequeño de un día para otro?" (síntomas improbables o absurdos); "¿tiene graves problemas con pensamientos acerca del suicidio?" (Síntomas flagrantes). La escala requiere menos de una hora para su aplicación.

Los resultados permiten la clasificación de las personas examinadas en las categorías de engaño definitivo, probable engaño y honestidad. La confiabilidad del instrumento es buena, con coeficientes de confiabilidad de consistencia interna que van de .66 a .92. Los estimados de confiabilidad intercalificadores son espléndidos, con un rango desde .89 a 1.00.

Aunque la validez de la SIRS puede analizarse en las líneas familiares de contenido, relacionada con el criterio y de constructo (y la prueba tiene un buen desempeño en estas áreas), la verdadera medida de su utilidad clínica tiene que ver con la capacidad de la prueba para discriminar a los simuladores conocidos o presuntos de los pacientes psiquiátricos y de los controles normales. Un estudio reciente indica que la prueba tiene un desempeño adecuado en este sentido (Gothard, Viglione, Meloy y Sherman, 1996). En una muestra combinada de 125 varones canalizados para valoración de competencia (que incluyó a 30 personas a las que se les pidió que fingieran simulación, siete individuos de los que existía una fuerte sospecha de simulación y 88 personas para las cuales ésta parecía poco probable), la SIRS tuvo una precisión general de 97.8% en la clasificación de los participantes en las categorías de simulación o no simulación.

Otros estudios revelan resultados prometedores, pero en ellos han participado principalmente muestras de personas de raza blanca y educadas (Rogers, Gillis, Dickens y Bagby, 1991; Rogers, Kropp, Bagby y Dickens, 1992). En contraste, la población del sistema de justicia criminal —el área en la que es más probable que se emplee la SIRS— es relativamente poco educada y existe una representación excesiva de minorías. En un estimado, más de 80% de la población urbana en cárceles en EUA es afroestadounidense (Dixon, 1995). Todavía no se ha respondido qué tan bien funcionará la SIRS en la detección de la simulación en esta subpoblación de tamaño considerable.

EVALUACIÓN DEL ESTADO MENTAL PARA LA DEFENSA DE INSANIA

En los juicios criminales, el acusado puede invocar una variedad de defensas que incluyen la incitación al delito, capacidad disminuida (p. ej., por subnormalidad mental), automatismo (p. ej., por sugestión hipnótica) y la defensa de insania. Cada vez que se invoca una defensa especial, se requiere una valoración del acusado en cuanto a su **estado mental al momento de cometer el delito** (MSO). En algunos tribunales, un psicólogo está calificado para ofrecer opiniones acerca del MSO de un acusado. En este caso se restringe el análisis a la defensa de insania, dado que es la doctrina más común que daría lugar a la necesidad de una valoración del MSO.

Casi todos estamos familiarizados con la defensa de insania, pero sólo algunas personas excepcionales comprenden sus disposiciones. En términos técnicos, la defensa de insania se conoce como **no culpable a causa de insania (NCCI)**. Con base en algunos juicios sensacionalistas y ampliamente publicitados, como el caso de John Hinckley quien intentó asesinar al presidente Ronald Reagan, en general el público lego ha concluido que los abogados cínicos emplean a menudo la defensa de insania para ayudar a sus clientes a evadir la responsabilidad de sus espantosos crímenes. Nada podría estar más lejano de la verdad. En realidad, el alegato de NCCI es muy respetado por los expertos en jurisprudencia, y se le invoca en menos de 1 de cada 1 000 procesos (Blau, 1984); y en esta mínima fracción de todos los casos criminales, la defensa tiene éxito en menos de una ocasión de cada cuatro (Melton y cols., 1998). La extendida creencia de que las

personas para las que se dictamina NCCI simplemente "se van" sin pagar sus crímenes, también es imprecisa: la mayoría recibe tratamiento hospitalario que dura varios años. Sus tasas de reincidencia son quizás inferiores (y seguramente no mayores) que las de los delincuentes convictos por ofensas similares (Melton *et al.*, 1998). Aunque ilegal en algunos estados, la defensa de insania ha mostrado una elasticidad notable —probablemente debido a que desempeña una función deseable en una sociedad moderna y compasiva.

Varias pruebas legales de este tipo han tenido una influencia significativa en EUA, incluyendo la regla M'Naughten, la regla de Durham, la regla del Código Penal Modelo y el veredicto de Culpable, Pero Mentalmente Enfermo (CPME); (Wrightsman y cols., 1994). Algunas jurisdicciones incluyen el impulso irresistible como complemento de la Regla M'Naughten. Algunos estados han abolido por completo la defensa de insania. Aquí se analizarán de manera breve las diferentes normas antes de comentar la función de las pruebas psicológicas en la determinación de la insania legal.

La **regla M'Naughten** es la más antigua y se deriva de un caso ocurrido en Inglaterra en 1843. Daniel M'Naughten estaba asediado por delirios paranoides acerca de que el primer ministro, Robert Peel, era parte de una conjura en su contra. M'Naughten acosó al primer ministro y en un caso de confusión de identidad le disparó a su secretario en la casa del No. 10 de la Calle Downing. A M'Naughten se le encontró no culpable a causa de insania, un veredicto que disparó el furor nacional. En respuesta a esta irritación, la reina Victoria le ordenó a los 15 jueces superiores de Inglaterra que aparecieran en la Cámara de los Lores y aclararan las recién creadas pautas sobre insania. La regla M'Naughten establece que:

> En todos los casos se le debe informar al jurado que todo hombre se supone cuerdo y que posee el suficiente grado de razón como para ser responsable de sus crímenes, hasta que se pruebe lo contrario a la satisfacción de dicho jurado; y que para establecer una defensa fundamentada en la insania debe probarse con toda claridad que, al momento de cometer el acto, el acusado se encontraba bajo tal defecto de razón, por enfermedad mental, que no conocía la naturaleza y calidad del acto que cometía; o, de conocerlo, no sabía que lo que estaba haciendo era incorrecto (citado en Wrightsman *et al.*, 1994).

Así, la regla M'Naughten "excusa" la conducta criminal si el acusado, como consecuencia de una "enfermedad mental", no sabía lo que estaba haciendo (p. ej., un individuo esquizofrénico paranoide que haya creído que le disparaba al mismo diablo) o que no sabía que lo que estaba haciendo era incorrecto (p. ej., una persona con retraso mental que haya creído que era aceptable dispararle a un mendigo fastidioso). Aproximadamente la mitad de los estados de EUA utilizan la Regla M'Naughten.

Algunas jurisdicciones aceptan el "impulso irresistible" como complemento de la regla M'Naughten. En general, se define como una respuesta conductual tan fuerte que el acusado no pudo resistirla a través de voluntad o razón. Pero ¿cuándo puede considerarse que un impulso es irresistible en oposición a simplemente no resistido? Esto ha resultado difícil de definir. Por obvias razones, los expertos legales no están nada felices con el concepto de impulso irresistible y en la actualidad se desvanece su utilización como parte del alegato de insania.

La **regla de Durham** se formuló en 1954 en el Tribunal de Apelaciones del Distrito de Columbia en el caso de **Durham contra EUA**. Insatisfecho con la regla M'Naughten, el juez David Bazelon propuso una nueva prueba, conocida como la regla de Durham, que regula la defensa de insania si el acto criminal fue "producto" de enfermedad o defecto mental. El propósito de la esta regla era darle a los profesionales de la salud mental mayor amplitud para presentar la información pertinente a la responsabilidad del acusado. Los eruditos legales le dieron la bienvenida a esta regla como un gran avance, pero en 1972 el mismo distrito judicial que la había formulado la eliminó.

La regla del Código Penal Modelo, propuesta por el Colegio de Abogados de EUA, reemplazó a la regla de Durham. Adoptada en

1972, la regla del **Código Penal Modelo** es la siguiente:

> Una persona no es responsable de conducta criminal si en el momento de presentar la conducta, y como resultado de enfermedad o defecto mental, carece de la capacidad sustancial ya sea para apreciar la criminalidad [incorrección] de su conducta o para someter su conducta a los requisitos de la ley (citado en Melton *et al.*, 1998).

Esta regla también contiene disposiciones que prohiben la inclusión de la psicopatía o de la personalidad antisocial dentro de la defensa de insania.

La regla del **Código Penal Modelo** difiere de la M'Naughten en tres sentidos importantes:

1. Al utilizar el término **apreciar**, reconoce los determinantes emocionales del acto criminal.
2. No requiere de una carencia total de reconocimiento de parte de los delincuentes en cuanto a la naturaleza de su conducta —sólo una falta de "capacidad sustancial".
3. Incluye tanto un elemento cognitivo como uno volitivo, lo cual hace que la incapacidad de los acusados para controlar sus actos sea un criterio independiente para la insania (Wrightsman *et al.*, 1994).

Cerca de 20 estados en EUA siguen ahora la regla del Código Penal Modelo o ligeras variaciones de la misma.

Un avance reciente en el alegato de insania es el veredicto de **Culpable, Pero Mentalmente Enfermo (CPME)**. Aproximadamente una cuarta parte de los estados le permiten a los jurados llegar a un veredicto de CPME en casos donde el acusado alega insania. Por lo común, en estos estados el juez instruye al jurado a entregar uno de los siguientes cuatro veredictos:

- culpable del delito
- inocente del delito
- no culpable a causa de insania
- culpable, pero mentalmente enfermo

La intención de la última alternativa es que un acusado que se encuentra CPME debería recibir el mismo fallo condenatorio que si se le hubiera encontrado culpable del delito, pero debe comenzar su sentencia en un hospital psiquiátrico. Después de la conclusión del tratamiento, el acusado cumple con la parte restante de su condena en una prisión.

Pero la intención del veredicto de CPME y su realidad son dos cosas diferentes. El apoyo inicial para este veredicto como variante humanitaria del alegato de insania se ha desvanecido en años recientes. Wrightsman *et al.* (1994) señalan que los jurados expresan confusión cuando se les pide hacer la difícil distinción entre enfermedad mental que resulta en insania (CPME) y aquella que no resulta en ello. Melton *et al.* (1998) encuentran pocas virtudes en el veredicto:

> ... el veredicto de CPME es conceptualmente equívoco, tiene un potencial significativo para engañar a la persona que trata de descubrir los hechos y no parece lograr su meta de reducir las exoneraciones por insania o prolongar el confinamiento de los delincuentes que están mentalmente enfermos y son peligrosos. La única meta que puede lograr es aliviar la ansiedad de jurados y jueces que de otra manera tendrían dificultad para decidir entre un veredicto de culpabilidad y uno de inocencia a causa de insania. Es dudoso que esta meta sea apropiada o digna de alabanza (p. 215).

Los estudios empíricos indican que los delincuentes con un veredicto de CPME pocas veces reciben tratamiento adecuado. Lo que es más, es posible que reciban condenas más duras que otros delincuentes a los que sólo se les encontró culpables (Callahan, McGreevy, Cirincione y Steadman, 1992). ¡De hecho, algunos acusados declarados CPME han sido sentenciados a morir!

Ahora que se ha introducido al lector a las variedades del alegato de insania, es posible revisar la función del psicólogo en la determinación de la insania legal. Un punto importante es que los psicólogos tienen toda la razón de ser cautos en ofrecer su opinión basada en una entrevista acerca del estado mental de una persona al momento de cometer un acto delictivo. Después de todo, por lo común, un delito ha ocurrido días, meses o incluso años antes, y es

posible que el cliente no sea capaz de ayudar en la reconstrucción precisa de los acontecimientos y estados mentales. En consecuencia, el testimonio psicológico acerca de la insania legal debe ser cauto y conservador. Los estudios de confiabilidad acerca de las valoraciones sobre insania también sugieren que es apropiada la precaución. En una reseña de siete estudios, Melton *et al.* (1998) determinaron que la concordancia intercalificadores (en cuanto a si un acusado sufría insania legal) iban desde un nivel bajo de 64% (entre los psiquiatras de la acusación y de la defensa) hasta un nivel elevado de 97% (entre psicólogos con entrenamiento forense que utilizaban los instrumentos estructurados que se analizan después).

A pesar de la controversia sobre la función de los psicólogos en la determinación del MSO, algunos expertos predicen un mayor papel de la evaluación psicológica en casos que implican el alegato de insania. En particular, las evaluaciones neuropsicológicas pueden proporcionar datos objetivos y válidos que ayuden a los tribunales a decidir los méritos de una defensa de insania. Las recientes resoluciones de los tribunales afirman que las conclusiones de una prueba neuropsicológica pueden utilizarse para mostrar que un acusado tiene un deterioro en su capacidad para elegir lo correcto y abstenerse de lo incorrecto (Blau, 1984; Heilbrun, 1992). Martell (1992) ha analizado la importancia de la evaluación neuropsicológica para el alegato de insania, como lo define el Código Penal Modelo, el cual define al acusado como no culpable a causa de insania si "carece de capacidad sustancial" para apreciar la criminalidad de su conducta. Los resultados de pruebas neuropsicológicas tienen importancia directa para esta cuestión.

Las escalas de calificación como las Escalas Rogers de Evaluación de la Responsabilidad Criminal (R-CRAS) pueden proporcionar también una base útil para valorar la responsabilidad criminal (Rogers, 1984, 1986). El examinador llena las R-CRAS inmediatamente después de revisar los registros clínicos, informes de investigación policiaca y la entrevista clínica final con el paciente-acusado. El instrumento consiste en criterios descriptivos claros para 25 reactivos que evalúan tanto factores psicológicos como situacionales. Ésos se califican con respecto al momento del delito en cinco escalas que miden las siguientes variables:

- Confiabilidad del paciente
- Organicidad
- Psicopatología
- Control cognitivo
- Control conductual

Los reactivos individuales en las R-CRAS se obtuvieron de la norma de insania del Código Penal Modelo (cuadro 10–10). Las confiabilidades intercalificadores para las escalas R-CRAS van desde .48 (para una subescala de Simulación) hasta 1.00 (para Organicidad). La validez de constructo se estableció al comparar la disposición tomada en 93 casos legales con los datos de las R-CRAS. Aunque al resultado legal lo determinan muchas variables además del estado psicológico de la persona al momento de cometer el delito, hubo 95% de concordancia en la determinación de cordura y 73% en la de insania.

Aunque los estudiosos reconocen que las R-CRAS son prometedoras, aún existe cierto escepticismo sano. Una preocupación es que las subescalas del instrumento representan un nivel **ordinal** de medición, mientras que se implica la existencia de cuantificación al nivel de **intervalo**. Otra preocupación es que los autores de la prueba afirman "cuantificar áreas de juicio que tienen una naturaleza lógica, intuitiva o ambas", lo cual conduce a un sentido falso de certidumbre científica (Melton *et al.*, 1998). Con toda seguridad, las R-CRAS desempeñan una función valiosa al ayudar a los clínicos a organizar su pensamiento y valoración. La utilidad de la decisión general —cuerdo contra insano— dependerá de investigación adicional sobre validez (Howell y Richards, 1989).

COMPETENCIA PARA ESTAR SUJETO A JUICIO

La Sexta Enmienda de la Constitución de EUA, aprobada en 1791, garantiza que todo ciudadano

Cuadro 10–10. Reactivos de muestra de las R-CRAS

10. Amnesia sobre el supuesto delito
(Esto se refiere a la evaluación de amnesia por parte del examinador, no necesariamente a aquella que informe el paciente.)
0) No existe información
1) Ninguna: recuerda todo el acontecimiento con detalles considerables
2) Ligera: de dudosa importancia. El paciente olvida detalles menores
3) Leve: el paciente recuerda la sustancia de lo que ocurrió, pero olvida muchos detalles menores
4) Moderada: el paciente ha olvidado una parte importante del supuesto delito, pero recuerda suficientes detalles como para creer que ha sucedido
5) Grave: El paciente sufre amnesia sobre la mayor parte del supuesto delito, pero recuerda suficientes detalles como para creer que ha sucedido
6) Extrema: el paciente tiene amnesia completa acerca de la totalidad del supuesto delito

11. Delirios al momento de ocurrir el supuesto delito
1) No existe información
2) Sospecha de delirios (p. ej., sustentado sólo en un autoinforme cuestionable).
3) Delirios confirmados, pero en realidad no relacionados con la comisión del supuesto delito
4) Delirios confirmados que contribuyeron al supuesto delito, pero que no constituyeron una fuerza predominante sobre la comisión del mismo
5) Delirios controladores confirmados, con base en los cuales se cometió el supuesto delito

Fuente: Adaptado y reproducido con autorización especial de la editorial, Psychological Assessment Resources, Inc., Odessa, FL 33556, de Rogers' Criminal Responsability Assessment Scales de Richard Rogers, Ph.D. Copyright 1984 de PAR, Inc. Se prohíbe mayor reproducción sin autorización de PAR, Inc.

acusado tiene el derecho a un juicio imparcial, expedito y público con el beneficio de una defensa. Si por cualquier razón, el acusado es incapaz de ejercer estos derechos constitucionales, entonces no puede llevarse a cabo un juicio apropiado. En términos específicos, si el acusado tiene un defecto, enfermedad o padecimiento mentales que lo hacen incapaz de comprender el proceso o de ayudar en su propia defensa, entonces se considera que éste es incompetente para estar sujeto a juicio. La Suprema Corte de EUA confirmó esta norma en el caso de **Dusky contra EUA** (1960) como "...si [el acusado] tiene la suficiente capacidad en el momento como para consultar con su abogado con un grado razonable de comprensión racional —y si tiene tanto una comprensión racional como objetiva acerca de las diligencias que se siguen en su contra". En la práctica, la **competencia para estar sujeto a juicio** se refiere a cuatro elementos y distinciones (Melton *et al.*, 1998):

1. La capacidad del acusado para comprender el proceso criminal, incluyendo el papel de los participantes en ese proceso.
2. La capacidad del acusado para funcionar en ese proceso, principalmente al ayudar a sus abogados en la preparación de una defensa.
3. La capacidad del acusado, al contrario de su disposición, para comunicarse con sus abogados y comprender los trámites.
4. El grado razonable de comprensión de parte del acusado, en contraposición a una comprensión perfecta o completa.

La mayor parte de los tribunales en EUA sigue esta norma que enfatiza el funcionamiento **actual** del acusado.

El juez que preside el caso puede solicitar una valoración psicológica o psiquiátrica para ayudar a determinar la competencia del acusado para someterse a juicio. Un informe reciente indica que cada año se realizan en EUA más de 25 000 valoraciones de competencia para estar sujeto a juicio (McDonald, Nussbaum y Bagby, 1992). Es importante enfatizar que los psicólogos, psiquiatras y otros profesionales de la salud mental sólo prestan ayuda en una audiencia de competencia al presentar sus opiniones expertas. Sólo el juez tiene el poder de tomar una determinación de competencia. Aunque

no existe un formato estándar para dicha determinación, la mayoría de los jueces solicitan que el psicólogo tome en consideración la mayor parte o los 11 factores que se citan en el cuadro 10–11.

La incompetencia para estar sujeto a juicio es algo totalmente independiente de la insania legal; estas dos cuestiones se juzgan con base en normas absolutamente diferentes. La insania legal tiene que ver con el momento de la comisión del acto delictivo, mientras la incompetencia implica una situación continua y actual; lo que es más, esta última no es sinónimo de enfermedad mental, aunque las dos pueden ocurrir al mismo tiempo. En caso de que el juez establezca que el acusado es incompetente, el juicio se pospone, en general, por un periodo de seis meses aproximadamente. En algunos casos, a las personas consideradas incompetentes se les interna en una institución mental para un tratamiento que tiene el propósito de restaurar su competencia, a fin de que después se pueda llevar a cabo un juicio. Los individuos acusados de delitos menos graves pueden recibir tratamiento como pacientes externos.

Cuadro 10–11. Factores considerados en la determinación de la competencia para estar sujeto a juicio

1. Compresión del acusado acerca de los cargos
2. Comprensión del acusado acerca de la naturaleza y amplitud de penas
3. Entendimiento del acusado acerca de la naturaleza del sistema de contrarios en el proceso legal
4. Capacidad del acusado para revelarle al abogado los hechos pertinentes que rodean a la supuesta ofensa
5. Capacidad del acusado para relacionarse con el abogado
6. Capacidad del acusado para ayudar al abogado en la planificación de la defensa
7. Capacidad del acusado para desafiar de manera realista a los testigos de la acusación
8. Capacidad del acusado para manifestar un comportamiento apropiado en el tribunal
9. Capacidad del acusado para atestiguar de manera pertinente
10. Motivación del acusado para ayudarse a sí mismo en el proceso legal
11. Capacidad del acusado para enfrentarse con el estrés del encarcelamiento previo al juicio

Fuente: Florida Rules of Criminal Procedure, citadas en Wrightsman, L. S., Nietzel, M. T. y Fortune, W. H. (1994). *Psychology and the legal system* (3a ed.) Pacific Grove, CA: Brooks/Cole.

Además de la información obtenida en la entrevista clínica, los resultados de pruebas psicológicas son componentes importantes de una valoración de competencia. Por ejemplo, un CI bajo puede constituir evidencia de incompetencia ante los ojos del tribunal. Aunque no existen pautas firmes, la mayor parte de los tribunales determinan que las personas con un déficit intelectual significativo —digamos un CI en el rango de retraso moderado o más bajo— son incompetentes para estar sujeto a juicio. De la misma manera, un patrón de resultados de prueba que indiquen una deficiencia neuropsicológica grave puede justificar una conclusión de incompetencia legal, aunque el CI del cliente se encuentre dentro del rango normal. Por ejemplo, un acusado que sufre de deficiencias graves de comprensión del lenguaje a causa de una embolia, puede considerarse incompetente para estar sujeto a juicio.

Se encuentran disponibles varias pruebas y procedimientos formales de detección que tienen la función de ayudar en la valoración de competencia. Estos instrumentos se resumen en el cuadro 10–12. Aquí se centra la atención en la *Competency Screening Test*, CST (Instrumento de Evaluación de la Competencia), que probablemente es la de más uso como auxiliar en la valoración de competencia.

Lipsitt, Lelos y McGarry (1971) desarrollaron la CST en un intento por formalizar la valoración de competencia. La CST consiste en 22 frases incompletas que se centran en los trámites del tribunal, relaciones abogado-cliente y procesos de pensamiento del acusado (figura 10–2). Cada respuesta se califica con 0 (deficiente), 1 (limítrofe pero no claramente apropiado) o 2 (claramente apropiado), según

Cuadro 10–12. Instrumentos de evaluación de competencia

***Competency Screening Test* (Instrumento de Evaluación de la Competencia) (Lipsitt, Lelos y McGarry, 1971)**

La CST es una prueba de completación de oraciones con 22 reactivos que tiene una guía objetiva de calificación. Algunos comentaristas consideran que los criterios de calificación son vagos y también se quejan de que la prueba tiene una deficiente validez aparente y débil validez de contenido.

***Competency Assessment Instrument* (*Laboratory for Community Psychiatry*, 1974)**

El CAI es una entrevista estructurada de 13 funciones relacionadas con el funcionamiento competente en un juicio. Una preocupación es que los examinadores emplean diferentes enfoques para aplicar la escala.

***Interdisciplinary Fitness Interview* (Entrevista Interdisciplinaria de aptitud) (Golding, Roesch y Schreiber, 1984)**

La IFI es una entrevista semiestructurada del acusado en cinco áreas legales y 11 categorías de psicopatología. En términos ideales, el abogado del acusado debe estar presente durante la aplicación.

***Fitness Interview Test* (Prueba de Entrevista de Aptitud) (McDonald, Nussbaum y Bagby, 1992)**

La FIT es una clasificación semiestructurada del acusado en cuanto a 24 aspectos legales y 14 áreas psiquiátricas. La FIT revela excelente confiabilidad intercalificadores y discrimina bien entre grupos de acusados clasificados como aptos o no para estar sujetos a juicio. Todavía está pendiente su revisión y se aconseja precaución en el empleo de la escala.

***Georgia Court Competency Test –Revised* (Prueba de Competencia del Tribunal de Georgia –Revisada) (Johnson y Mullett, 1987)**

La GCCT-R es una prueba oral con 21 reactivos sobre comprensión de los procedimientos en un tribunal, conocimiento del cargo, conocimiento de las posibles penas y capacidad para comunicarse de manera racional. Algunos revisores advierten que la prueba toma una muestra inadecuada del dominio de las capacidades relacionadas con la competencia.

muestras de normas de calificación. En un estudio de validación cruzada basado en 50 residentes varones de una unidad estatal forense, la CST predijo de manera correcta 82% de las recomendaciones de competencia presentadas por el equipo forense encargado de determinar la competencia o incompetencia legal. Se encontró que la CST estimaba de manera precisa la competencia, pero pronosticaba en exceso la incompetencia (Nottingham y Mattson, 1981). En general, la CST e instrumentos similares son útiles al inicio de una valoración de competencia, pero no deben constituir el único medio de evaluación. Otras pruebas de competencia se revisan en Bagby, Nicholson, Rogers y Nussbaum (1992); Melton *et al.* (1998) y Nicholson, Robertson, Johnson y Jensen (1988).

Una grave preocupación en las valoraciones de competencia es si el cliente está simulando. Después de todo, demorar la fecha del juicio durante un largo tiempo proporciona un motivo fuerte para parecer incompetente. Los clínicos tienen una variedad de métodos y pruebas (descritos anteriormente) para identificar a clientes que podrían estar simulando; aun así, el proceso para la valoración de competencia no es inmune, como lo indican casos tan sonados como el del hombre de Connecticut que evadió una acusación por asesinato (Prensa Asociada, 30 de junio de 1998). Presuntamente, este individuo había asesinado a su novia anterior y al novio actual de ésta con una pistola, para después dispararse en la cabeza. Sufrió daño cerebral y parálisis parcial, y **cuatro** psiquiatras lo declararon incompetente para estar sujeto a juicio en cuatro audiencias separadas. Los psiquiatras afirmaron que era incapaz de comunicarse de manera efectiva con su abogado. Se anuló una orden de la corte que requería que este hombre se sometiera a valoraciones anuales de competencia, lo cual permitió que se le pasara por alto y quedara libre. Nueve años después se le encontró asistiendo a una universidad como estudiante de los cursos de preparación para la carrera de medicina y con un promedio de calificaciones entre 8 y 9. Ejemplos como este constituyen una razón para

1. El abogado le dijo a Bill que _____
2. Cuando vaya al tribunal, el abogado habrá de _____
3. Jack sintió que el juez _____
4. Cuando acusaron a Phil de un delito, él _____
5. Cuando me prepare a ir al tribunal con mi abogado _____
6. Si el jurado me encuentra culpable, yo _____
7. La manera en que se decide un juicio _____
8. Cuando se presentó al jurado la evidencia del caso de George _____
9. Cuando el abugado interrogó a su cliente en el tribunal, éste dijo _____
10. Si Jack tuviera que presentar su propio caso, él _____
11. Cada vez que el fiscal de distrito me hacía una pregunta, yo _____
12. Mientras escuchaba a los testigos atestiguar en mi contra, yo _____
13. Cuando el testigo en contra de Harry dio evidencia falsa, él _____
14. Cuando Bob no estuvo de acuerdo con su abogado defensor, él _____
15. Cuando se me acusó formalmente del delito, pensé para mí _____
16. Si el abogado de Ed le sugiere que se declare culpable, él _____
17. Lo que más le preocupa a Fred acerca de su abogado _____
18. Cuando dicen que un hombre es inocente hasta que se pruebe lo contrario _____
19. Cuando pienso en que me envíe a prisión, yo _____
20. Cuando Phil piensa en aquello por lo que lo han acusado, él _____
21. Cuando el jurado escuche mi caso, ellos habrán de _____
22. Si tuviera oprotunidad de hablar con el juez, yo _____

2. Cuando vaya al tribunal, el abogado habrá de
 a) Criterio legal: capacidad para cooperar, comunicarse y relacionarse con su defensor.
 b) Criterio psicológico: capacidad para relacionarse o confiar.

Calificación 2:
Ejemplos: "defenderme"
"estar allí para ayudarme"
"hacer lo mejor posible para obtener una sentencia leve"
"representarme"
"presentar mi caso"

Calificación 1:
Ejemplos: "estar allí"
"pedir que el juicio se posponga"
"pedirme que declare"

Calificación 0:
Ejemplos: "encerrarme"
"mantener la boca cerrada"
"acusarme"

Figura 10–2. *Competency Screening Test* (Instrumento de Evaluación de la Competencia).
Fuente: Reimpreso con autorización de Lipsitt, P. D. (1970), *Competency Screening Test.* Boston: Competency to Stand Trial and Mental Illness Project.

mostrar humildad y cautela cuando los psicólogos abordan valoraciones de competencia.*

* Otra explicación (cínica) es que incluso las personas incompetentes con daño cerebral grave recibirán calificaciones de 10 o 9 si asisten a la universidad. Quizá la inflación de calificaciones ha ido demasiado lejos.

PREDICCIÓN DE LA VIOLENCIA Y EVALUACIÓN DEL RIESGO

En ocasiones, los psicólogos fungen como asesores durante la fase de sentencia de un juicio para determinar si un acusado convicto repre-

senta un peligro para sus compañeros en prisión o para los miembros de la comunidad. Esta información puede ser fundamental para determinar la duración de una sentencia o la asignación del acusado (p. ej., a centros de seguridad mínima, intermedia o máxima). Las decisiones sobre libertad condicional también pueden depender de una determinación de peligrosidad. En casos extremos, el testimonio en una corte de los profesionales de la salud mental puede influir en una decisión para invocar la pena de muerte.

Por desgracia, aunque muchos investigadores han intentado diseñar medidas psicológicas de la peligrosidad, no se ha desarrollado ninguna prueba o procedimiento psicológico que proporcione una predicción precisa y de largo alcance de la violencia (Blau, 1984; Horowitz y Willging, 1984; Melton *et al.*, 1998). De hecho, ni siquiera es posible **confirmar posteriormente**, con cierto grado de precisión, la conducta violenta a partir sólo de pruebas psicológicas.

¿Por qué es tan difícil predecir con precisión la conducta violenta? Melton *et al.* (1998) citan cuatro factores que contribuyen a este desafío:

1. **Variabilidad en la definición legal.** El resultado a predecir se define de diversas maneras como violencia, riesgo o peligrosidad, y los criterios de pertinencia legal varían de un ambiente a otro. Algunos estatutos requieren de pruebas de un acto explícito dentro de un periodo particular, mientras otros podrían aceptar un juicio clínico de "tendencias explosivas" como prueba de violencia.
2. **Complejidad de la literatura.** La literatura de investigación que examina la relación entre factores antecedentes y reincidencia en conducta violenta es mucha; no obstante, es "tanto abrumadora como inconexa", de modo que incluso los profesionales más escrupulosos encuentran difícil extraer algún significado de ella.
3. **Errores y parcialidad en los juicios.** Los profesionales de la salud mental están propensos a cometer errores cognitivos sutiles en sus valoraciones de los acusados. Por ejemplo, un problema es la tendencia a considerar a la peligrosidad únicamente como un rasgo cuando, de hecho, su aparición siempre es producto de una interacción con factores ambientales (p. ej., disponibilidad de un arma). Otro problema es que los clínicos permiten que la notoriedad de un delito afecte de manera desproporcionada su juicio.
4. **Consecuencias políticas para el profesional.** La predicción excesiva de violencia es más segura para los clínicos que la limitada. Si se libera a un acusado con base en la predicción de un clínico y después dicho acusado comete un delito violento, el profesional es vulnerable a una acción legal debido a liberación negligente. En los casos limítrofes, existe un incentivo muy poderoso para inclinarse en la dirección de una conclusión de "peligrosidad".

Un estudio reciente de Menzies, Webster, McMain, Staley y Scaglione (1994) ilustra las dificultades en la predicción de la conducta violenta. Estos investigadores utilizaron un *Dangerous Behavior Rating Scheme*, DBRS (Esquema de Calificación de la Conducta Peligrosa) de 15 reactivos para predecir conducta violenta en un seguimiento a seis años con 162 acusados enviados a evaluación de peligrosidad. El DBRS consistió en 15 reactivos, en una escala Likert de siete puntos, que medían atributos de la personalidad (p. ej., hostilidad, emocionalidad, capacidad de empatía), factores ambientales (p. ej., apoyo, estrés), factores situacionales (p. ej., peligrosidad aumentada con uso de alcohol) y reactivos generales (p. ej., peligro hacia otras personas en el presente). Además, se utilizó una variedad de índices de predicción basados en las calificaciones de enfermeras, psiquiatras, trabajadores sociales y otros profesionales para pronosticar una variedad de medidas de resultado en intervalos que iban de 1 a 6 años. La mayor parte de las correlaciones se colocó en el orden de .2 y muy pocos elementos de predicción lograron llegar al límite superior de .40. La correlación más alta fue de .53, pero tenía que ver con el caso

restringido de predicción de violencia a corto plazo en individuos hospitalizados en un pabellón psiquiátrico. En general, los resultados fueron tan desalentadores que los investigadores aconsejaron en contra "de la participación clínica o psicométrica en la identificación de sujetos potencialmente violentos en ambientes clínicos o correccionales".

Una inquietud particular en la predicción de violencia es la tasa muy elevada de falsos positivos que suele observarse en estudios controlados. Esto significa que un número importante de personas para las que se predice violencia posteriormente no revelan ninguna conducta de este tipo. Gardner, Lidz, Mulvey y Shaw (1996) le pidieron a clínicos en una sala de urgencias psiquiátricas que calificaran su preocupación (de 0, ninguna preocupación, a 5, gran preocupación) acerca de que 784 pacientes se vieran involucrados en actos violentos a lo largo de los siguientes seis meses. Se combinaron las calificaciones del clínico y del psiquiatra de guardia (un total de 10 puntos posibles) para producir un estimado más estable. Un total de 327 pacientes recibieron puntuaciones totales de 6 o mayores, lo cual indica un grado importante de preocupación clínica. De este grupo se encontró que ¡49% estaban completamente libres de conducta violenta durante los siguientes seis meses! Una fórmula estadística simple que utilizó información sobre antecedentes resultó significativamente más precisa en la realización de predicciones, pero aun así produjo una tasa de 42% de falsos positivos. Ambos métodos —el clínico y el estadístico— también produjeron niveles altos de falsos negativos (pacientes de los que no existía sospecha de violencia futura y que subsecuentemente participaron en actos de ese tipo).

Aunque la predicción de **largo alcance** sobre peligrosidad por medio de pruebas psicológicas o escalas de calificación ha resultado difícil, la evaluación de **corto alcance** sobre el riesgo con delincuentes violentos ha logrado un éxito moderado. Con un análisis del historial previo de los delincuentes violentos, Hall (1987) ha demostrado que pueden utilizarse las siguientes variables para obtener una probabilidad de violencia durante los tres meses posteriores a la valoración forense:

- Ocurrencia reciente del acto o actos violentos anteriores.
- Número de actos graves de violencia ocurridos antes.
- Abuso de sustancias en el curso del mes anterior.
- Amenaza de rompimiento o rompimiento real de una relación amorosa.
- Problemas laborales que conducen a acción disciplinaria o despido.
- Oportunidad para la violencia; por ejemplo, acceso a una pistola.

Al cuantificar los factores anteriores, el examinador llega a una probabilidad de que la violencia continúe en el futuro a corto plazo (Hall, 1987). Esta estrategia es útil principalmente con personas evaluadas que tienen un historial de conducta violenta, de modo que tiene una aplicabilidad limitada en el campo forense. No obstante, para ciertos individuos que satisfacen la mayor parte o todos los criterios listados antes, es más frecuente que se llegue a conclusiones correctas en la predicción de conducta violenta. El enfoque que se destaca aquí demuestra una vez más el poder de los datos biográficos en la predicción de la conducta.

VALORACIÓN DE CUSTODIA DE LOS NIÑOS EN EL DIVORCIO

Es posible que los psicólogos testifiquen en las disputas sobre custodia infantil que surgen después del divorcio. De hecho, dichas disputas son poco comunes —en el 90% de los casos de divorcio, ambos padres llegan a un acuerdo sobre las disposiciones de custodia, sin recurrir a la intervención legal (Melton *et al.*, 1998). La función del consultor consiste en ofrecer opiniones expertas sobre temas como idoneidad de los padres o intereses primordiales del niño. En general, estas opiniones se basan en evaluaciones del niño (o niños) y de ambos padres y puede incluir pruebas psicológicas.

Por desgracia, en pocas ocasiones el testimonio que se basa en pruebas psicológicas proporciona una base útil para ayudar al juez a tomar la decisión salomónica que se requiere en una disputa de custodia (Blau, 1984). La debilidad esencial de las pruebas psicológicas en este contexto es que el vínculo entre las conclusiones de prueba y la paternidad eficaz es débil o no existe. La técnica Rorschach y otras pruebas de personalidad no se diseñaron para evaluar la relación de los padres con sus hijos y, por tanto, son irrelevantes en gran medida para las cuestiones reales que aparecen en los casos de custodia infantil.

Desgraciadamente, los clínicos han sido lentos para darse cuenta de sus limitaciones en el área forense en relación con las disputas de custodia infantil. Como resultado "probablemente no existe pregunta forense en la que los profesionales de la salud mental se hayan extralimitado de manera tan común y atroz" (Melton *et al.*, 1998, p. 484). Es comprensible que los profesionales legales se muestren escépticos acerca del valor del testimonio psicológico en casos de custodia infantil.

La *American Psychological Association* ha reconocido la tensión que existe entre la psicología y el sistema legal al publicar las pautas para las valoraciones de custodia infantil en los trámites de divorcio (APA, 1994a). Estas pautas se refieren a preocupaciones acerca del "mal uso de la influencia de los psicólogos" en los trámites de custodia y ofrecen principios para la práctica en esta área cada vez más compleja. Las pautas especifican que los intereses primordiales del niño son lo más importante. Especifican además que se requiere de entrenamiento especializado, incluyendo familiaridad con las normas y procedimientos legales aplicables y conocimiento de las leyes que gobiernan el divorcio y la adjudicación de custodia en la jurisdicción local. Las pautas también dictan una completa neutralidad de parte del psicólogo:

> El psicólogo, de manera equilibrada e imparcial, informa y aconseja a la corte y a los posibles custodios del niño acerca de los factores psicológicos relevantes relacionados con las cuestiones de custodia. El psicólogo debe ser imparcial, sin importar si lo ha contratado la corte o una de las partes dentro de los trámites (APA, 1994, p. 678).

Esta última pauta tiene especial importancia para la profesión, en vista de que promueve la dignidad en la práctica de la psicología. Sin embargo, la imparcialidad es una cualidad difícil de mantener, en la medida en que surgen presiones sutiles cuando un abogado de una de las partes contrata los servicios de un psicólogo para la valoración de custodia.

En respuesta a la ausencia de una base de datos científica, se han propuesto varios instrumentos experimentales de evaluación para los procedimientos forenses de custodia infantil, incluyendo escalas de calificación que intentan estimar el potencial para la crianza infantil eficaz. Una de tales pruebas es la *Ackerman-Schoendorf Scales for Parent Evaluation of Custody* (Escalas Ackerman-Schoendorf de Valoración de los Padres para Custodia; Ackerman y Schoendorf, 1992). La ASPECT es, en esencia, una batería de pruebas estándar (Rorschach, MMPI-2, WAIS-R, WRAT-R para los padres y la Prueba del Dibujo de una Familia, la CAT y una medida de CI para el niño) que se utilizan junto con cuestionarios de preguntas abiertas, entrevistas, observaciones y registros del tribunal (en caso necesario), como base de datos a partir de la cual el profesional llena un inventario de 56 reactivos con respuestas de sí-no (se llenan inventarios independientes para el padre y la madre). Los resultados de los 56 reactivos de sí-no producen tres subpuntuaciones y una calificación total, el Índice de Custodia Paterna (ICP) para cada uno de los padres. Todas las calificaciones se informan como puntuaciones *T* con media de 50 y desviación estándar de 10. El ICP tiene el propósito de identificar cuál de los padres es más eficaz —y qué tanto. Una diferencia en puntuación *T* de 10 a 15 puntos se considera significativa, de 16 a 20 puntos son muy significativas de más de 20 puntos son notables. Cuando ambos padres tienen puntuaciones en el ICP superiores a 60, se considera probable que cualquiera de ellos sería un progenitor eficaz. Si ninguno de ellos lo es (puntuaciones de 40 o menores), el

ICP tiene el propósito de reflejar también esa conclusión.

Para la puntuación total del ICP, la confiabilidad intercalificadores va de .92 a .96 y se considera adecuada. Los datos de confiabilidad para las tres subescalas, son:

- Observacional, o la apariencia y presentación del progenitor
- Social o de interacción del progenitor con otras personas, incluyendo al niño
- Cognitiva-emocional, o sobre el funcionamiento psicológico y mental del progenitor

Son sustancialmente más débiles, lo cual indica que los profesionales deben depender sólo de la puntuación del ICP. Se utilizó una amplia muestra de estandarización (N = 200), pero estas familias eran predominantemente de raza blanca (97%), de modo que es posible que el ASPECT no sea apropiado para utilizarse con familias de minorías.

La validez de la puntuación del ICP se evaluó al comparar las recomendaciones de la ASPECT con las decisiones de los jueces en casos de custodia. En aquellos casos donde existió una diferencia significativa entre las puntuaciones de la ASPECT para la madre y el padre (10 puntos o más), la prueba mostró 93% de concordancia con las decisiones de los jueces. Aun así, los revisores han recomendado precaución en el empleo de la ASPECT por varias razones. Una preocupación es la escasa investigación independiente en publicaciones de consulta que se relacione con la validez del instrumento. Melton (1995) señala que el estudio de validación se basó en las mismas familias utilizadas para desarrollar el instrumento, lo cual puede conducir a tasas infladas de acierto. Otra preocupación es que no se ha demostrado que algunos de los reactivos de la escala (p. ej., si el CI de uno de los padres es inferior en más de cinco puntos al del niño; cuál de los padres se coloca junto a éste en su dibujo) indiquen competencia del progenitor o resultado para el niño en estudios sobre divorcio. Heinze y Grisso (1996) concluyen que la ASPECT requiere de mayores datos normativos, de confiabilidad y de validez antes de que los profesionales puedan colocar su confianza en la batería de evaluación.

Un problema con la ASPECT es el gran tamaño de la batería completa, que incluye pruebas extensas con ambos padres y con todos sus hijos, así como entrevistas y análisis de abundante (¿excesiva?) información de prueba. Quizá pudiera obtenerse el mismo tipo de información de evaluación de un modo que consuma menos tiempo. Éste es el enfoque que toman Gordon y Peek (1989), quienes desarrollaron el *Custody Quotient™* (Cociente de Custodia), una elaborada escala de calificación basada en reactivos que produce una calificación de nueve puntos para cada uno de los siguientes factores relacionados con los padres:

Necesidades emocionales del niño ahora y en el futuro.
Necesidades físicas del niño ahora y en el futuro.
Ausencia de peligros emocionales o físicos para el niño.
Buenas habilidades de cuidado parental.
Asistencia del progenitor.
Planificación para el niño.
Estabilidad en el hogar.
Cuidado previo.
Actos u omisiones.
Valores.

El instrumento también produce un cociente general de custodia o CC (media de 100), diseñado para estimar la competencia de los padres. Los autores informan sobre algunos datos empíricos acerca de su instrumento, incluyendo estudios de concordancia intercalificadores basados en valoración de entrevistas grabadas en vídeo. En la mayor parte de las escalas, los dos autores se colocaron a una distancia de dos puntos entre sí (en escalas de calificación de nueve puntos) en 90 a 100% de las ocasiones. La validez del instrumento se ha evaluado, en parte, a través del análisis factorial confirmatorio que revela un "factor de buena paternidad en general" como se predijo. Un estudio inicial de validación reveló que los padres con un CC menor

a 100 en pocas ocasiones recibieron la custodia del niño (excepto en casos donde ambos padres recibieron calificaciones menores a 100). Gordon y Peek (1989) reconocen que su instrumento requiere investigación adicional.

Otra prueba para la valoración de la custodia infantil es el *Parent-Child Relationship Inventory*, PCRI (Inventario de Relaciones Padres-Hijo, Gerard, 1993). El PCRI es un inventario de autoinforme con 78 reactivos, adecuado para madres y padres de niños de 3 a 15 años de edad. El inventario incluye siete escalas:

- Apoyo del progenitor.
- Satisfacción.
- Participación.
- Comunicación.
- Establecimiento de límites.
- Autonomía.
- Orientación del rol.

Las normas se basan en una muestra no clínica de más de 1 000 padres y madres en todo EUA. Como la mayor parte de los instrumentos relacionados con la valoración de la custodia infantil se requiere de más investigación para determinar el valor del PCRI para este tipo de evaluaciones. Una preocupación especial es que los reactivos de prueba son relativamente transparentes, de modo que los padres pueden sesgar a nivel consciente las conclusiones de la prueba en una dirección favorable.

Bricklin (1995) ha diseñado cuatro pruebas sencillas que han resultado útiles en valoraciones de custodia. Dos de ellas evalúan habilidad y dedicación en los cuidados parentales para la madre y el padre por separado. En términos específicos, el *Parent Awareness Skills Survey* (Examen de Consciencia de Habilidades del Progenitor) evalúa la conciencia sobre las habilidades de crianza infantil, mientras que el *Parent Perception of Child Profile* (Perfil de Percepción Paterna acerca del Niño) valora un interés genuino en el niño. En este caso se centrará la atención sobre los dos instrumentos restantes que examinan la percepción del niño acerca de sus padres: Las *Bricklin Perceptual Scales*, BPS (Escalas Perceptuales Bricklin) y la *Perception-of-Relationships Test*, PORT (Prueba de Percepción de Relaciones).

Las BPS consisten en 64 preguntas que se le hacen a cada niño (de cuatro años de edad en adelante) implicado en una disputa de custodia. Hay 32 preguntas que tienen que ver con la percepción del niño acerca de su madre y 32 se relacionan con percepciones sobre el padre. Las preguntas se encuentran dentro de cuatro áreas diseñadas para identificar los intereses primordiales del niño: capacidad, apoyo, consistencia y posesión de rasgos admirables. Lo siguientes son ejemplos de los tipos de preguntas que se realizan:

- **Capacidad**: "Si tienes problemas con una boleta de calificaciones de la escuela, ¿cuánto sería capaz de ayudarte tu mamá?
- **Apoyo**: "Cuando te sientes realmente preocupado, ¿cuánto te ayuda tu papá a tranquilizarte?"
- **Consistencia:** "Si trataras de seguir dormido en un día de escuela ¿con cuánta frecuencia te obligaría tu mamá a levantarte a la hora de siempre?"
- **Posesión de rasgos admirables:** "Si tuvieras una mascota, ¿qué tan bien cuidaría de ella tu papá si te fueras de campamento durante una semana?"

Las 32 preguntas sobre mamá y las 32 sobre papá son idénticas; se encuentran ampliamente separadas en el cuestionario, a fin de evitar comparaciones inmediatas y obvias. Las preguntas se hacen primero de manera verbal y después se repiten con un fraseo un tanto diferente. Por ejemplo, el examinador sostiene una lámina de prueba frente al niño; "si éste (señalando al extremo marcado con "Muy Bien") es tu papá si te ayudara muy bien a tranquilizarte y éste (señalando al extremo marcado con "No Tan Bien") es tu papá sin ayudarte no tan bien a tranquilizarte, ¿dónde estaría tu papá en esta línea?" El niño responde utilizando un punzón para hacer un orificio en la lámina. Cada una contiene una rejilla de calificación en el lado opuesto para una calificación objetiva e inmediata.

Las calificaciones de la BPS consisten en una comparación reactivo por reactivo de las

calificaciones para mamá y papá. Se identifica el Progenitor de Elección (PDE) en términos de cuál de ambos, la madre o el padre, obtiene la mayor calificación en la mayor parte de los reactivos:

> Se supone que este progenitor es el más capacitado para operar en beneficio de los "intereses primordiales" del niño en la variedad más amplia de situaciones como se les mide con la BPS (Bricklin, 1995, p. 77).

Por supuesto, este resultado es sólo un punto de la información que utiliza el examinador para llegar a una recomendación sobre custodia.

La BPS posee varias cualidades positivas en comparación con otros enfoques para la valoración de custodia. Primero, la prueba evita pedirle al niño que elija de manera directa entre sus padres, lo cual reduce al mínimo la culpa. También evalúa importantes áreas de atención y cuidado que se suponen importantes para el desarrollo sano del niño. Lo que es más, al centrarse en las percepciones del niño acerca del cuidado que le brindan sus padres, en contraposición a las percepciones del examinador acerca de ello, el instrumento proporciona una voz para aquellos que están en mayor riesgo en una decisión sobre custodia. Bricklin (1995) señala que la BPS no necesariamente tiene el propósito de ser una medida directa de la capacidad de los padres, sino que se le diseñó para identificar a la persona cuyo estilo es congruente con la capacidad del niño para aprovechar y beneficiarse de la orientación parental.

Aunque la evidencia acerca de la confiabilidad de la BPS no se encuentra disponible en el trabajo de Bricklin (1995), las conclusiones iniciales sobre validez la apoyan. En términos específicos, la BPS produjo 90% de tasa de concordancia con las elecciones de progenitor hechas por profesionales de salud mental, quienes tuvieron acceso a datos independientes sobre antecedentes clínicos y familiares obtenidos durante varios años. Como nota precautoria debe decirse que, como ocurre con tantos instrumentos en el campo de la valoración sobre custodia, existen pocos (o nulos) estudios empíricos sobre la BPS publicados en publicaciones de consulta. Heize y Grisso (1996) ofrecen la siguiente crítica:

> Una cuestión relacionada es que las Escalas Bricklin intentan responder las dudas legales sobre preferencias de custodia sin evaluación directa de las capacidades y deficiencias funcionales de los padres, incluyendo la manera en que éstas se relacionan con las necesidades de los niños. Además, las escalas atienden sólo de manera muy superficial a la cuestión de dos padres con deficiencias significativas. Así, debido a que las percepciones del niño acerca de ambos padres se comparan sólo entre sí y no con aquéllas sobre otros padres, es posible que el clínico no obtenga mucha información concerniente al nivel de funcionamiento parental de un progenitor específico como lo percibe el niño (p. 301).

Es claro que sería deseable mayor investigación de parte de fuentes independientes.

Otra prueba inventada por Bricklin (1995) y que también se basa en el niño es la *Perception-of-Relationships Test*, PORT (Prueba de Percepción de Relaciones). La PORT es una prueba de dibujo proyectivo que puede calificarse para determinar al progenitor de elección. Se pide al niño que lleve a cabo siete conjuntos de dibujos, comenzando con uno de cada uno de sus padres. La prueba prosigue entonces con otros dibujos que incluyen una autorrepresentación en la misma hoja que el dibujo de la mamá y después en la del papá. Se utiliza un complejo sistema de calificación para identificar al progenitor con el que el niño busca la "cercanía" psicológica. Se considera que ello revela la percepción del niño acerca de cuál de sus padres le presta más apoyo y es más capaz de responder a sus necesidades. En una serie de siete estudios de validez a través de varios decenios (Bricklin, 1992), la PORT solía tener 90% de precisión en la predicción del progenitor de elección como lo designó un juez de un tribunal o clínicos independientes.

Una aplicación especializada para la PORT que menciona Bricklin (1995) es la detección del maltrato físico y del abuso sexual. Los criterios de calificación que indican maltrato o abuso parecen haberse derivado del saber psicoanalítico. Por ejemplo, se considera que

todos los siguientes son signos de maltrato o abuso:

- Notable incremento en la distancia entre la figura de sí mismo y la del progenitor que ha cometido maltrato o abuso en comparación con la del otro progenitor que no maltrata o abusa de él
- Líneas onduladas o interrumpidas en el área corporal en la que se ha sufrido maltrato o abuso (p. ej., área genital, pecho)
- Distorsión de una mano en el dibujo del progenitor que ha cometido maltrato o abuso

Por desgracia, el apoyo empírico para estas hipótesis es limitado, como se analiza con amplitud en el Tema 13B, Técnicas proyectivas. Debido a que los alegatos de abuso sexual han sido tan perjudiciales en los trámites sobre custodia, los examinadores deben ejercer gran cuidado al especular sobre tales cuestiones, en especial si la evidencia consiste únicamente en pocos dibujos de la figura humana que se consideran sospechosos.

Para cerrar esta sección se le recuerda al lector que el testimonio experto en **valoraciones sobre custodia** infantil deben moderarse con una enorme dosis de humildad de parte del psicólogo examinador. Determinar los intereses primordiales del niño es casi siempre una tarea difícil y espinosa. La ausencia de pruebas y métodos adecuadamente validados para esta tarea garantiza que la valoración de custodia continuará siendo una de las responsabilidades más desafiantes en toda la práctica psicológica.

LESIONES PERSONALES Y TESTIMONIO RELACIONADO

Las lesiones personales como las causadas por un accidente de automóvil son, a menudo, una fuente de demandas en búsqueda de compensación económica. En las demandas por **lesiones personales**, los abogados pueden contratar a psicólogos para que testifiquen en cuanto a las consecuencias crónicas del estrés traumático o del daño cerebral provocado. Por ejemplo, un neuropsicólogo clínico podría aplicar una batería completa de pruebas (capítulo 9, Evaluación neuropsicológica y geriátrica) y después atestiguar en cuanto a las implicaciones funcionales a largo plazo de un daño cerebral conocido.

En general, un psicólogo consultor que rinde testimonio ante un tribunal se enfrentará con normas sumamente elevadas para ejercicio profesional. Ya se ha mencionado la norma Frye, que dispone que el testimonio debe basarse en pruebas y procedimientos que hayan "obtenido aceptación general" en el campo. Así, una prueba o procedimiento que es relevante o útil en la práctica clínica cotidiana —pero que no ha recibido total aceptación en el campo— podría recibirse con escepticismo dentro de los tribunales. Incluso un juez podría resolver que el testimonio es inadmisible si se basa en pruebas o procedimientos con una validación frágil. Peor aún, el juez puede aceptar dicho testimonio, lo cual deja al testigo experto a merced de la crítica y ridiculización de los abogados de la parte contraria. Con estas preocupaciones en mente, Heilbrun (1992) ha publicado pautas para la práctica de la evaluación forense que se resumen a continuación:

- Las pruebas deben estar disponibles a nivel comercial y estar bien documentadas, como en el *Mental Measurements Yearbook*.
- La confiabilidad debe estar bien documentada y ser de .8 o mayor, excepto en circunstancias poco comunes.
- Las pruebas deben relacionarse con la cuestión legal o con un constructo psicológico subyacente a la misma.
- Deben utilizarse los métodos estándar de aplicación, con condiciones ideales de prueba.
- Las pruebas deben haberse validado en una población apropiada que se relacione con el individuo al que se evalúa.
- Cuando sea posible, los profesionales deberán utilizar pruebas objetivas con fórmulas actuariales para la interpretación.
- Los profesionales deben ser cuidadosos al verificar la existencia de simulación, defen-

sividad y otras razones que obliguen a pasar por alto o ignorar los datos de prueba.

Pautas adicionales pueden encontrarse en las *Specialty Guidelines for Forensic Psychologists* (*Comittee on Ethical Guidelines for Forensic Psychologists*, 1991).

Cada vez es más frecuente que los tribunales hayan estado dispuestos a otorgar compensaciones por daños mentales además de daños físicos. Este tipo de lesiones se conoce como "trauma psíquico", "angustia emocional" o "perjuicio emocional". La evaluación del daño emocional dependerá, en cierta medida, de los resultados de una prueba psicológica (en particular, pruebas de personalidad), pero la evaluación requiere gran habilidad clínica que incluya "antecedentes longitudinales del daño, su tratamiento y los intentos de rehabilitación, incluyendo la motivación del demandante para recuperarse" (Melton *et al.*, 1998, p. 381). Una vez más puede observarse que la cuestión de la simulación persigue a la mayor parte de las formas de evaluación forense.

INTERPRETACIÓN DE PRUEBA DE POLÍGRAFO

Aunque por lo general los resultados de polígrafo no se aceptan en los casos judiciales, cerca de 20 estados de EUA permiten el testimonio si ambas partes en un proceso están de acuerdo en que se utilice la evidencia. Cuando menos dos estados (Massachusetts y Nuevo México) permiten el testimonio aun si la parte contraria se opone (Wrightsman *et al.*, 1994). En casos en los que se emplea la prueba de polígrafo, es posible pedir a los psicólogos que ofrezcan su testimonio experto. A continuación se proporciona una breve revisión de la prueba de polígrafo.

William Marston, un estudiante de Hugo Munsterberg en la Universidad de Harvard, fue quien desarrolló el polígrafo alrededor de 1917. El rudimentario enfoque de Marston consistía en medir la presión sanguínea sistólica de un sospechoso durante el interrogatorio. Este investigador consideraba que la presión sanguínea sistólica se elevaría de manera gradual si la persona examinada intentaba engañar al examinador. Su polémica técnica fue el procedimiento que se puso en duda en el sobresaliente caso de **Frye contra EUA** (1923). El lector recordará que la decisión del caso Frye ayudó a definir los requisitos solicitados de un testigo experto, al definirlo como la persona que utiliza un método que "... ha obtenido aceptación general en el campo particular al que pertenece". En el caso Frye, el tribunal no aceptó el testimonio de Marston, dado que su enfoque de "detector de mentiras" no había recibido amplia aceptación.

El **polígrafo** moderno es sustancialmente más complejo que su rudimentario predecesor creado por Marston. En una prueba moderna de polígrafo se conectan varios monitores al cuerpo de la persona examinada: un brazalete para medir presión, un monitor de frecuencia cardiaca, un aro flexible alrededor del pecho para registrar la respiración y conductos eléctricos en los dedos para detectar los cambios en actividad eléctrica de la piel. El polígrafo registra las respuestas físicas que ocurren al momento, las cuales incluyen cambios en respiración, pulso, presión sanguínea y transpiración. Los cambios en esta última no se miden de manera directa, sino por medio de un par de electrodos que estiman los cambios en conductividad eléctrica en la superficie de la piel. Incluso un ligero incremento en la humedad debida a la transpiración facilita la conductividad de una corriente eléctrica sobre la superficie de la piel. La palabra polígrafo se traduce de modo literal como "muchas escrituras". Las respuestas fisiológicas se presentan en un conjunto de líneas fluctuantes dibujadas en tinta sobre un rollo de papel en movimiento continuo.

La justificación para el empleo del polígrafo como detector de mentiras se deriva de la observación de que muchas personas reaccionan con un aumento en la excitación fisiológica cuando dicen una mentira. Al menos en teoría, las respuestas veraces se acompañan de líneas relativamente planas, mientras que se supone que una mentira causa fluctuaciones

significativas y factibles de detección en frecuencia cardiaca, transpiración y quizás también en otras medidas. Así, un procedimiento común en la prueba de polígrafo consiste en comparar las respuestas a una pregunta neutra o control ("¿hoy es jueves?") con respuestas a una pregunta relacionada ("¿robó usted al Primer Banco Interestatal el pasado viernes?").

Aunque es común llamar al polígrafo "detector de mentiras", esta denominación coloquial es sustancialmente imprecisa, debida en gran parte, a los propios expertos en su uso. En consecuencia, en nuestros días rodea al polígrafo una vasta mitología. De hecho, este aparato detecta las respuestas fisiológicas, no las mentiras. Una respuesta fisiológica distintiva registrada por un polígrafo puede o no ser señal de una mentira —ésta es una cuestión empírica abierta a investigación.

El polígrafo no "suena" al momento de decir una supuesta mentira; es un examinador quien debe interpretar el patrón de respuestas fisiológicas. En esto reside una de las limitaciones sustanciales de dicho instrumento: se requiere una gran cantidad de interpretación basada en juicios para determinar cuáles "señales" son significativas y cuáles son coincidencias. Por desgracia, los juicios "expertos" de los técnicos experimentados en el polígrafo no se sostienen muy bien ante las pruebas empíricas en el mundo real (Carroll, 1988; Lykken, 1981, 1987). En un estudio de campo acerca de la precisión de la interpretación del polígrafo, Kleinmuntz y Szucko (1984) presentaron lo que quizás es el retrato menos halagador acerca de la precisión de los expertos en polígrafo. Los autores recolectaron lecturas de polígrafo de 50 personas de quienes se probó posteriormente que eran ladrones y de otras 50 de quienes había sospechas sobre los mismos robos y que después fueron exoneradas. Además, se añadieron 20 casos no verificados, como de "amortiguación" o "relleno", pero que no se incluyeron en los análisis.

Las pruebas de polígrafo se presentaron a seis expertos sumamente experimentados en el uso del instrumento, y se les dijo que la mitad de la muestra era culpable y la otra mitad inocente. Los resultados del estudio consistieron en el cómputo de las identificaciones correctas e incorrectas por parte de los intérpretes. El lector reconocerá que los intérpretes estarían en lo correcto en 50% de las ocasiones sólo por adivinar. También debe señalarse que eran posibles varias categorías de clasificación: positivos válidos (personas culpables identificadas como tales), negativos válidos (personas inocentes identificadas como tales), falsos positivos (personas inocentes identificadas como culpables) y falsos negativos (personas culpables identificadas como inocentes). Kleinmuntz y Szucko (1984) también convirtieron las lecturas de polígrafo en datos digitales con propósitos de un análisis funcional discriminativo. En un análisis de este tipo, se utilizan factores estadísticos puramente objetivos para clasificar a los participantes con la mayor precisión posible.

Los resultados del estudio fueron inesperadamente desalentadores. En promedio, los interpretes estuvieron en lo correcto en 69% de las ocasiones, sólo 19% por arriba de la mera adivinación. El análisis funcional discriminativo tuvo un desempeño un tanto mejor, llegando a 73%. El mejor interprete estuvo en lo correcto sólo 76% de las veces. Las tasas de error en este cuidadoso estudio fueron alarmantemente elevadas, en especial en el caso de los falsos positivos —personas inocentes identificadas como culpables. Incluso el mejor intérprete de polígrafo clasificó a 18%, o 9 de las 50 personas inocentes, como culpables.

Blinkhorn (1988) resume de la siguiente manera una perspectiva psicométricamente informada acerca del polígrafo:

> El empleo del polígrafo como detector de mentiras está muy lejos de satisfacer los estándares aceptables para las pruebas psicológicas. En esencia, no tiene estandarización; es inconsistente a nivel interno; la recalificación de las gráficas es poco confiable; no se dispone de ningún tipo de información de confiabilidad para los examinandos; produce un número desproporcionado de resultados positivos falsos; no se ha investigado su impacto adverso en subgrupos sociales y étnicos; implica la medición de aspectos sensibles del temperamento (p. 39).

Otros análisis son más favorables, pero no tanto como para que los expertos legales se apresuren a promover el polígrafo como evidencia en el tribunal. Saxe, Dougherty y Cross (1985) examinaron 10 estudios de campo acerca de la precisión del polígrafo. En éstos, la tasa promedio de clasificación correcta de personas inocentes fue de 89% (rango de 77 a 100%), mientras que aquélla para individuos culpables fue de sólo 82% (rango de 47 a 100%). Recientemente, un importante estudio de campo sobre la prueba de polígrafo por parte del Servicio Secreto de EUA concluyó que los examinadores identificaron de manera correcta a 96% de los individuos veraces y a 95% de los sospechosos que mintieron (Raskin, 1989). Elaad, Ginton y Jungman (1992) también analizaron la precisión del polígrafo en investigaciones delictivas reales. Recopilaron los registros de polígrafo de pares de personas acusadas del mismo delito (p. ej., robo de una grabadora de vídeo de una base militar). En total, utilizaron los registros de 40 personas inocentes y 40 culpables, de quienes posteriormente se estableció la verdad a través de confesión voluntaria. Al utilizar reglas objetivas de decisión, se clasificó de manera correcta a 94% de los individuos inocentes y se identificó de modo adecuado a 76% de los culpables. En conjunto, todos estos estudios brindan apoyo al empleo del polígrafo en las etapas de investigación de un caso delictivo. No obstante, las tasas de error son todavía demasiado elevadas como para que se acepte como evidencia en un tribunal (Honts y Perry, 1992).

En años recientes, el polígrafo parece haber tenido un retorno limitado. Por ejemplo, el estado de Texas controla a los delincuentes sexuales con pruebas de polígrafo, lo cual contrapone los derechos civiles con la seguridad del público (Prensa Asociada, 16 de mayo de 1997). En un caso, un varón de 33 años que estaba en libertad condicional por abusar de un niño pequeño, "falló" una prueba de detector de mentiras concerniente al contacto sexual con niños. Cuando se le confrontó con ello, confesó haber tenido actividad sexual con un joven de 15 años de edad y después enfrentó una sentencia de prisión por violar la libertad condicional. Cada vez es más frecuente que los sistemas judiciales dependan del detector de mentiras para vigilar a los delincuentes sexuales (50 000 tan sólo en el estado de Texas), lo cual hace surgir preguntas legales y éticas difíciles. Una preocupación es el problema de los falsos positivos, en los que se tiene sospecha de un delito sexual en el caso de una persona inocente con base en las crestas observadas en las líneas continuas de tinta de una prueba de polígrafo. En una encuesta con 195 psicólogos de la *Society for Psychophysiological Research*, la mayoría de los entrevistados respondieron que la detección de mentiras por este medio no es teóricamente sólida, que medidas contrarias que se aprenden con facilidad pueden vencer a una prueba de mentiras y que los resultados de prueba no deberían admitirse en los tribunales (Iacono y Lykken, 1997).

CONTROVERSIA ACERCA DEL PSICÓLOGO COMO TESTIGO EXPERTO

A partir de los temas anteriores debería ser evidente que el testimonio experto de los psicólogos puede alterar la vida de las personas. Como observan Faust y Ziskin (1988): "Dependiendo de la opinión del experto, un individuo puede terminar recluido en una institución mental, recibir fuertes recompensas monetarias, obtener la custodia de un niño o perder la vida". Estos autores también señalan que los clínicos participan cuando menos en un millón de casos legales cada año, de modo que el impacto potencial del testimonio experto es trascendental. Es adecuado concluir esta sección sobre cuestiones legales y judiciales con una breve mención acerca de la polémica que rodea al psicólogo como testigo experto.

En un libro en tres volúmenes intitulado *Coping With Psychiatric and Psychological Testimony*, Ziskin y Faust (1988) proporcionan una perspectiva completamente escéptica acerca del clínico en los tribunales. Estos autores

resumen sus principales conclusiones en un artículo más breve (Faust y Ziskin, 1988). Su crítica reside en dos afirmaciones:

1. Los testigos expertos en psicología (y psiquiatría) no pueden responder a las preguntas forenses con una precisión razonable.
2. Estos expertos no ayudan al juez y al jurado a llegar a conclusiones más exactas de lo que sería posible de otro modo.

La primera afirmación se fundamenta en gran medida en la supuesta incapacidad de los clínicos para proporcionar diagnósticos psiquiátricos correctos. El argumento va más o menos así: si los clínicos no pueden desempeñar de manera precisa una tarea tan fundamental como la clasificación o diagnóstico de los pacientes, ¿cómo sería posible que realizaran el trabajo más complejo de determinar el estado previo, actual o futuro de la persona bajo escrutinio? Así, los problemas con la confiabilidad del diagnóstico se utilizan para ilustrar la dificultad más general de lograr concordancia entre los clínicos acerca de las descripciones o predicciones del estado pasado, actual o futuro de la persona examinada. Para apuntalar su postura, Ziskin y Faust (1988) citan numerosos estudios que demuestran la falta de confiabilidad del diagnóstico psiquiátrico, incluyendo algunos análisis que muestran que la tasa de desacuerdo para categorías diagnósticas específicas es igual o superior a la tasa de acuerdo.

El segundo argumento se basa en la conjunción de resultados de investigación que sugieren que los clínicos profesionales no realizan, de hecho, juicios clínicos más correctos que los legos (Ziskin y Faust, 1988). Por ejemplo, Faust y Ziskin (1988) citan un estudio clásico de Goldberg (1959) en el que secretarias de oficina tuvieron un desempeño tan adecuado como el de los psicólogos profesionales en distinguir las producciones visomotoras de personas normales contra las de aquéllos con daño cerebral en el Test Guestáltico Visomotor de Bender, una prueba de copia de diseños. En el estudio de Goldberg (1959), la mitad de los dibujos en la prueba de Bender eran de pacientes con enfermedad cerebral orgánica confirmada, mientras la otra pertenecía a individuos normales; por tanto, la precisión diagnóstica aleatoria era de 50%. Se pedía a los participantes en el estudio que realizaran un juicio para los dibujos de cada uno de los individuos con dos alternativas: normal contra daño cerebral. En promedio, el personal de psicólogos con doctorado clasificó de manera correcta a 65% de los protocolos, los estudiantes de posgrado en psicología tuvieron 70% de éxito y las secretarias alcanzaron 67% de precisión. Ziskin y Faust (1988) citan muchos otros estudios que ilustran el mismo tema general de inexistencia de diferencias en la precisión de los juicios clínicos entre clínicos profesionales y legos.

En defensa del testimonio experto, varios autores han criticado el análisis de Faust y Ziskin (1988) debido a su academicismo combativo y a su falta de equilibrio (Brodsky, 1989; Fowler y Matarazzo, 1988; Heilbrun, 1992; Matarazzo, 1990). Por ejemplo, en su revisión de los estudios acerca de la precisión del diagnóstico, Matarazzo (1983, 1990) concluyó que los resultados indican magnitudes de confiabilidad que van de adecuadas a muy adecuadas; este autor cita muchos estudios positivos que no se mencionan en el trabajo de Faust y Ziskin (1988). Lo que es más, Fowler y Matarazzo censuran el énfasis sesgado sobre la precisión diagnóstica:

> La clasificación diagnóstica por sí sola es, en pocas o ningunas ocasiones, la base sobre la cual se toman determinaciones legales. Cuestiones legales como la insania, discapacidad y competencia se basan más en la comprensión del juez o del jurado acerca de la conducta y capacidad de un individuo para funcionar en situaciones específicas que en el diagnóstico concreto que se le asigna al mismo.

Coincidimos en que Faust y Ziskin (1988) revisaron la bibliografía deliberadamente unilateral y que ignora las conclusiones que apoyan a la psicología forense. Por ejemplo, considérese un detalle del estudio de Goldberg (1959) —un detalle que, no por coincidencia, Faust y Ziskin (1988) pasan por alto. Además de utilizar a psicólogos, profesionales en entrenamiento

y secretarias como jueces, Goldberg (1959) le pidió a un renombrado experto en la prueba Bender que clasificara los protocolos. El experto logró las mejores tasas de clasificación entre todas: un impresionante 83% correcto.* Con base en una revisión de estudios sobre juicio clínico, Garb (1992) concluye que los profesionales de la salud mental (y en especial los psicólogos) son más precisos en sus juicios que los legos; es decir, **tienen** algo significativo que ofrecer a jueces y jurados en casos forenses. Es claro que no puede negarse la falibilidad del clínico individual, pero esta cuestión no justifica la eliminación de los expertos en los tribunales.

RESUMEN

1. La práctica de la psicología confluye con el sistema legal en diversas áreas que incluyen la valoración de simulación, evaluación del estado mental (para la defensa de insania), evaluación de competencia para estar sujeto a juicio, predicción de violencia y valoración de custodia infantil.
2. Para fungir como testigo experto, un psicólogo debe ser un experto calificado en opinión de la corte. En general, el testimonio se restringe a técnicas que han estado disponibles por un tiempo prolongado a fin de tener antecedentes de aceptación general (**Frye contra EUA**, 1923).
3. En evaluaciones forenses, el profesional debe tener fuertes sospechas de simulación, pero debe llegar a tal conclusión de manera cauta debido a las consecuencias para el cliente. Herramientas como la Entrevista Estructurada de Síntomas Informados pueden ayudar en la evaluación de la simulación.
4. Los expertos en jurisprudencia tienen amplio respeto por la defensa de insania, y este argumento se invoca en menos de 1 de cada 1 000 juicios. En la actualidad, la mayor parte de los estados en EUA siguen la regla M'Naughten o el Código Penal Modelo como pruebas legales de insania. Algunas jurisdicciones aceptan el "impulso irresistible" como complemento.
5. Según la regla M'Naughten, puede considerarse que una persona no es culpable debido a insania si "al momento de cometer el acto, la parte acusada se encontraba bajo tal defecto de razón, de enfermedad de la mente, que no conocía la naturaleza y calidad del acto que cometía; o, de conocerlo, no sabía que era incorrecto".
6. El Código Penal Modelo propone que "una persona no es responsable de conducta criminal si en el momento de ocurrir dicha conducta, como resultado de enfermedad o defecto mental, carece de la capacidad sustancial ya sea para apreciar la criminalidad (incorrección) de su conducta o para someterla a los requisitos de la ley".
7. Una adición reciente a la jurisprudencia legal es el veredicto de Culpable Pero Mentalmente Enfermo. Su empleo tiene varios defectos que incluyen la confusión por parte de los jurados y las sentencias más duras que si sólo se encontrara culpable a la persona.
8. Las escalas de calificación como las Escalas Rogers de Evaluación de Responsabilidad Criminal proporcionan una base útil para valorar el estado mental del acusado al momento de la ofensa (MSO). No obstante, los analistas sugieren precaución en la aplicación de ésta y otras escalas similares debido a que podrían conducir a un falso sentido de certidumbre científica acerca de cuestiones inherentemente intuitivas y de juicio.

* El experto fue Max Hutt, quien después escribió The Hutt Adaptation of the Bender Gestalt Test (1977).

9. Es frecuente que los tribunales pidan a los psicólogos que ayuden a determinar si un acusado es competente para estar sujeto a juicio. En general, la competencia implica que el acusado es capaz de comprender los cargos en su contra, puede cooperar con el abogado de su defensa y dar una buena presentación de sí mismo durante el juicio.
10. La predicción de violencia y evaluación del riesgo es otra tarea vinculada con tribunales que en ocasiones se solicita de los psicólogos. Por desgracia, no se ha desarrollado ninguna prueba o procedimiento psicométrico que proporcione una predicción de largo alcance de la violencia. En contraste, la predicción de corto alcance de la violencia o riesgo ha encontrado mayor éxito.
11. Áreas adicionales de evaluación en cuestiones delicadas incluyen las disputas sobre custodia infantil en las que, en pocas ocasiones, los resultados de pruebas psicológicas son útiles; y los casos de daños personales, en los que las pruebas psicológicas pueden representar una función importante para documentar los efectos del daño.
12. Se han propuesto varios instrumentos y escalas de evaluación para proporcionar un medio objetivo para realizar recomendaciones sobre custodia infantil. A pesar de que proporcionan información útil, estas herramientas pueden prometer más de lo que cumplen, desde el punto de vista de validez psicométrica.
13. Un polígrafo detecta las respuestas fisiológicas, como los cambios eléctricos de la piel ante ciertas preguntas de interrogatorio, y las representa como líneas de tinta sobre papel. La investigación apoya al empleo del polígrafo en las etapas de investigación de un caso delictivo, pero las tasas de error aún son demasiado elevadas como para utilizarlo en juicios.
14. Los escépticos discuten que los testimonios de expertos en psicología (y psiquiatría) no pueden responder a las preguntas forenses con precisión razonable y no ayudan al juez y ni al jurado a llegar a conclusiones más correctas de lo que sería posible de otra manera. Estas opiniones parecen basarse en revisiones unilaterales de la investigación.

TÉRMINOS Y CONCEPTOS CLAVE

Testigo experto p. 452
Simulación p. 454
Estado Mental al Momento de cometer el Delito (MSO) p. 455
No culpable a causa de insania (NCCI) p. 455
Regla M'Naughten p. 456
Regla de Durham p. 456
Código Penal Modelo p. 457
Culpable Pero Mentalmente Enfermo (CPME) p. 457
Competencia para estar sujeto a juicio p. 459
Valoraciones sobre custodia p. 469
Lesiones personales p. 469
Polígrafo p. 470

Evaluación industrial y organizacional

CAPÍTULO 11

TEMA 11A
Evaluación y selección de personal

Esquema para la evaluación y selección de personal
Función de las pruebas en la selección de personal
Datos autobiográficos
Entrevista de trabajo
Pruebas de capacidad cognitiva
Pruebas de personalidad y temperamento
Pruebas de integridad de lápiz y papel
Estructura y medición de habilidades psicomotrices
Muestra de trabajo y ejercicios situacionales
Ejemplo de caso 11–1. Asignación de carrera en Canadá
Resumen
Términos y conceptos clave

La psicología industrial y organizacional (psicología I/O) es una subespecialidad de la psicología que trata sobre la conducta en situaciones de trabajo (Dunnette y Hough, 1990, 1991, 1992). En su sentido más amplio, la psicología I/O incluye diversas aplicaciones en negocios, publicidad y el ejército. Por ejemplo, es común que las empresas consulten a los psicólogos I/O para ayudar a diseñar y valorar los procedimientos de contratación; los negocios pueden solicitarles que evalúen la eficacia de la publicidad; y los líderes militares dependen en gran medida de ellos para el examen y colocación de reclutas. Las pruebas psicológicas al servicio de la toma de decisiones acerca del personal (contratación, colocación, ascensos y valoración) son, por tanto, un importante punto de esta profesión. Por supuesto, los especialistas

en psicología I/O tienen amplias habilidades y con frecuencia manejan muchas responsabilidades corporativas que no se mencionan antes. No obstante, no puede negarse el papel central de la evaluación en su profesión.

El propósito de este capítulo es revisar las numerosas maneras en que los psicólogos I/O pueden utilizar las pruebas y los procedimientos de evaluación para los propósitos antes mencionados. También se introducirá al lector en algunas de las cuestiones y controversias que pueden surgir en la selección y evaluación de los empleados. En el tema 11A, Evaluación y selección de personal, se analizan las prácticas de prueba para la selección de empleados y se introducen los instrumentos pertinentes. En el tema 11B, Estimación del desempeño en el trabajo, se atiende el complicado problema de la valoración del desempeño. También se revisan allí los aspectos legales y los efectos de los reglamentos gubernamentales sobre la evaluación industrial.

ESQUEMA PARA LA EVALUACIÓN Y SELECCIÓN DE PERSONAL

ASPECTOS CLAVE EN LA EVALUACIÓN DE PERSONAL

Es importante tener un esquema para comprender el papel intrincado y multifacético del psicólogo I/O en la evaluación y selección de personal. Un elemento crucial —que con frecuencia se pasa por alto en los análisis de los enfoques sobre pruebas y evaluación— es que los psicólogos I/O deben ser más que meros especialistas en pruebas. Este elemento es de particular importancia en la selección de personal, la cual requiere de mucho más que el simple conocimiento sobre las pruebas que podrían utilizarse. A fin de utilizar de manera prudente las pruebas para esta labor, el psicólogo I/O debe estar versado en diversas cuestiones complejas. Por ejemplo, el conocimiento detallado de la teoría de decisión, que se analiza en el tema 4A, Conceptos básicos de validez, es esencial para la utilización eficaz de las pruebas en la selección de personal. Elementos de la teoría de decisión, como las tablas Taylor-Russell, proporcionan una base para determinar la proporción esperada de solicitudes exitosas seleccionadas con una prueba. Un psicólogo I/O sería muy torpe en embarcarse en un nuevo programa de selección de empleados sin considerar las tablas Taylor-Russell y la teoría de decisión. Además, el asesor prudente debería tener un conocimiento bien fundamentado acerca de los conceptos sobre sesgo de prueba que se analizan en el tema 7B, Sesgo de una prueba y otras controversias. Por razones legales y éticas, los negocios desdeñan el empleo de pruebas tendenciosas. Por lo común, es responsabilidad del psicólogo la selección o desarrollo de pruebas que no tengan sesgo, incluyendo los criterios objetivos mediante los cuales se valora éste. De la misma manera, el especialista en personal debe comprender las cuestiones legales relacionadas con las pruebas de empleo, las cuales se analizan al final de este capítulo.

En el cuadro 11–1 se esbozan las muchas preocupaciones de los psicólogos I/O en la selección de personal, el cual proporciona una sinopsis de los principales problemas relacionados con las pruebas y la evaluación. Esta lista de aspectos clave no pretende ser exhaustiva, pero sí trasmite la amplitud de preocupaciones que se enfrentan en la aplicación de pruebas para la selección de personal. El lector observará que un área de conocimiento experto que el psicólogo I/O necesita es el análisis de puestos, el cual consiste de la identificación de los criterios para el desempeño eficaz en el trabajo. Un análisis de puestos concienzudo puede proporcionar los elementos esenciales para la selección de personal. Por tal razón, se analizan los enfoques acerca del mismo antes de volver a otras cuestiones dentro de la selección de personal.

ANÁLISIS DE PUESTOS

En grandes empresas y pequeños negocios por igual, los elementos esenciales del éxito de la organización son los trabajos individuales que

Cuadro 11–1. Recopilación de cuestiones clave en las pruebas y evaluación para selección de personal

Análisis de puestos: ¿Cuáles son los criterios específicos para el desempeño eficaz en el puesto?

Pruebas y evaluaciones: ¿Las estrategias y procedimientos de selección poseen una relación demostrada con el desempeño eficaz en el puesto?

Puntuaciones de corte: ¿Cuál es la proporción esperada de solicitantes exitosos seleccionados con una prueba de validez conocida?

Efectividad de costos: Si se incluyen los costos de las pruebas y selección, ¿cuáles procedimientos de evaluación producen el mayor beneficio general?

Sesgo de la prueba: ¿Las pruebas y los procedimientos de evaluación evidencian sesgo en contra de una o más minorías?

Pautas legales: ¿Las pruebas y los procedimientos de selección satisfacen las pautas federales sobre exámenes imparciales de empleo?

Estudios de validez: ¿Cuál es la validez en curso del programa de selección de personal?

llevan a cabo empleados específicos de todos niveles —desde la gerencia hasta el recluta neófito. La manera en que se definen los puestos es crucial para la dirección y crecimiento de la organización. Cascio (1987) ha descrito las preguntas que pueden hacerse cuando se examinan las organizaciones nuevas o existentes desde la perspectiva de lo puestos de trabajo:

> ¿Cuántos puestos deberán cubrirse y cuál será la naturaleza de éstos? ¿Qué capacidades, habilidades y características personales se requerirán de los empleados? ¿Cuántos individuos se reclutarán? ¿Qué factores (personales, sociales y técnicos) deberán tenerse en cuenta en la selección de estos individuos? ¿Cómo se les debe entrenar y qué criterios se utilizarán para medir qué tan bien se han desempeñado en su trabajo?

Antes de que pueda responderse a estas preguntas, los administradores deben definir los puestos en cuestión y después identificar las habilidades y conductas del empleado que son necesarias para desempeñar cada labor. Este proceso, conocido como análisis de puestos, cae de manera apropiada dentro del campo de acción del psicólogo I/O. En este caso se revisarán algunos de los problemas de evaluación que surgen del análisis de puestos y se revisarán algunos cuestionarios estandarizados que se utilizan para este propósito (Harvey, 1991).

En términos generales, el **análisis de puestos** consiste de la definición de un puesto en términos de las conductas necesarias para cumplir con él. Éste incluye dos componentes principales: descripción y especificación del puesto. La descripción del puesto identifica las características físicas y ambientales del trabajo a realizar, mientras que la especificación del mismo detalla las características personales necesarias para realizar dicho trabajo. Por ejemplo, una descripción de un puesto secretarial podría señalar que la persona debe ser capaz de manejar ocasionalmente quejas telefónicas, mientras que la especificación correspondiente del mismo podría listar como esencial la tolerancia hacia personas difíciles. Otro ejemplo: una descripción del puesto de oficial de policía podría especificar que la persona debe ser capaz de "Cargar a un individuo que ha sido arrestado y que no puede o se niega a caminar, al tomar a la persona por debajo de los brazos, soportando su peso, y transportarla a la patrulla", mientras que la correspondiente especificación del puesto podría referirse a un nivel mínimo de fuerza y vigor físicos. En este caso se está adelantando a los hechos, pero incluso sería posible poner en términos operaciones el concepto de "fuerza y vigor" como aquello que se requiere para arrastrar a un maniquí de 55 kilos a una distancia de 15 metros en un minuto (Arvey, Landon, Nutting y Maxwell, 1992).

La justificación principal para el análisis de puestos es que ayuda a proporcionar una base válida para tomar decisiones sobre el personal. Esto no sólo es deseable desde un punto de vista ético, sino que también satisface las normas legales decretadas por los tribunales y organismos reguladores. En el caso de **Albemarle contra Moody** (1975), la Suprema Corte de EUA determinó que debe incluirse el análisis de puestos en los estudios de validación que pretendan

demostrar una relación entre un instrumento de selección y el desempeño en un puesto. También, las *Uniform Guidelines on Employee Selection* (1978) requieren del análisis de puestos como un componente de los estudios sobre validez. Como especifica Cascio (1987), justificaciones adicionales para el análisis de puestos incluyen: planificación de los recursos de la organización, entrenamiento y desarrollo del personal, seguridad y mejoría de los métodos de trabajo e investigación sobre personal.

Pueden utilizarse numerosos métodos para realizar un análisis de puestos y un examen amplio nos llevaría demasiado lejos. En pocas palabras, Cascio (1987) lista los siguientes enfoques:

- Observación directa de los titulares del puesto.
- Entrevista estructurada de los trabajadores.
- Recolección de incidentes críticos por parte de los supervisores.
- Listas de verificación de deberes y habilidades.
- Cuestionarios.

Otra fuente útil de información para el análisis de puestos es el *Dictionary of Occupational Titles* o *DOT* (*U.S. Department of Labor*, 1991, 1993). El *DOT* proporciona información estandarizada sobre la ocupación para dar apoyo a las actividades de colocación en un empleo. Por ejemplo, la especificación del *DOT* para **psicómetra** consiste de lo siguiente:

> 045.067-018 PSICÓMETRA (profes. y sim.)
>
> Aplica, califica e interpreta pruebas de inteligencia, aprovechamiento y otras pruebas psicológicas para proporcionar información a maestros, orientadores, estudiantes y otras partes interesadas específicas. Aplica pruebas de lápiz y papel o utiliza equipo de prueba, bajo condiciones estándar. Cronometra, examina y registra los resultados. Interpreta los resultados de prueba bajo normas estándar y limitaciones de prueba en términos de validez y confiabilidad (U.S. Department of Labor, 1991).

El *DOT* también proporciona información laboral pertinente a la preparación educativa, demandas físicas del trabajo y condiciones ambientales que se enfrentan por lo general. Por ejemplo, el puesto de psicómetra requiere de 2 a 4 años de educación superior y se define como una ocupación sedentaria con frecuente necesidad de alcanzar, manejar, manipular y hablar. La buena audición y visión cercana también se listan como características físicas necesarias para el puesto (*U.S. Department of Labor*, 1993). El *DOT* es una importante fuente de información para el análisis de puestos, en particular para los puestos tradicionales que están bien definidos. De manera reciente, el *U.S. Department of Labor* (Ministerio del Trabajo de EUA) comenzó una ambiciosa iniciativa para analizar casi todos los puestos dentro de la economía, según el contenido o actividades laborales requeridas (Peterson, Mumford, Borman y otros, 1995). La base de datos resultante se denomina O**Net* y promete ser muy valiosa para establecer correspondencias entre personas y empleos.

Pero no todos los puestos están definidos claramente, lo cual significa que puede ser necesario que el psicólogo I/O lleve a cabo un análisis formal de puestos. Esto es más difícil de lo que podría parecer. Morgeson y Campion (1997) han resumido los escollos de un análisis de puestos preciso, los cuales incluyen factores sociales como la conformidad en los comités de análisis de puestos y problemas con la sobrecarga de información que se deriva de la complejidad de la tarea. En este caso se centra la atención sobre los cuestionarios para el análisis de puestos, dado que éstos hacen surgir varias cuestiones interesantes en cuanto a evaluación.

Los cuestionarios estructurados y cuantificables para el análisis de puestos no se volvieron populares sino hasta que el *Position Analysis Questionnaire* (PAQ; cuestionario de análisis de puestos) apareció a inicios del decenio de 1970 (McCormick, Mecham y Jeanneret, 1972). Este instrumento introdujo un concepto novedoso "orientado al trabajador" acerca del análisis de puestos, el cual permitía que los psicólogos I/O realizaran comparaciones significativas entre diferentes empleos. Aún popular hoy en día, el PAQ consiste de 194

reactivos o elementos del puesto que provienen de cinco categorías:

- **Obtención de información:** ¿Cómo y de dónde obtiene el trabajador la información necesaria para el puesto?
- **Procesos mentales:** ¿Qué tipos de razonamiento, planificación y toma de decisiones requiere el puesto?
- **Producto laboral:** ¿Cuáles son las actividades físicas que se llevan a cabo y las herramientas o instrumentos que utiliza el trabajador?
- **Relaciones personales:** ¿Qué tipos de relaciones con otras personas son inherentes al puesto?
- **Contexto del puesto:** ¿Cuáles son los contextos físicos y especiales en los que se lleva a cabo el trabajo?

Algunos de los elementos individuales del puesto simplemente se marcan con "sí" o "no", mientras que otros reactivos se califican en una escala apropiada como importancia, tiempo o dificultad. Por ejemplo, en una subsección, el analista de puestos califica (en una escala de 1 a 5) cinco aspectos de la comunicación oral (aconsejar, negociar, persuadir, ordenar y entrevistar) según su importancia para un puesto específico.

El PAQ tiene una confiabilidad respetable cuando lo utilizan personas entrenadas para el análisis de puestos. La confiabilidad intercalificadores se coloca por lo común alrededor del .80 para el instrumento en general (McCormick y cols., 1972; Mecham, McCormick y Jeanneret, 1977). Su validez se ha establecido de manera menos adecuada, aunque Cascio (1987) ha señalado que el género del analista, el nivel de interés de los titulares del puesto o la cantidad de información proporcionada acerca del puesto no afectan los resultados del PAQ. No obstante, el instrumento tiene limitaciones importantes, incluyendo el elevado nivel de lectura que se requiere (posgrado universitario para algunos reactivos) e idoneidad limitada para puestos profesionales y gerenciales. Quizás una limitación más grave es que el énfasis en los aspectos conductuales del trabajo puede pasar por alto importantes diferencias en tareas entre puestos sumamente diferentes. Por ejemplo, Arvey y Begalla (1975) han señalado que el perfil de un oficial de policía es bastante similar al de una ama de casa. Por ejemplo, ambos puestos pueden implicar resolución de conflictos —pero en contextos sumamente diferentes.

En respuesta a las deficiencias percibidas en el PAQ, Cornelius y Hakel (1978) desarrollaron el *Job Element Inventory*, JEI (Inventario de elementos del puesto). Similar en formato al PAQ, el JEI consiste de 153 reactivos con un nivel de lectura de educación secundaria. La innovación más significativa es que el instrumento tiene la suficiente sencillez como para que lo llenen los titulares del puesto. Dado que no se requiere de personas entrenadas en análisis de puestos, el JEI es sumamente eficiente en cuanto a costos. En un estudio analítico factorial del JEI con 2 000 titulares de puestos en la Guardia Costera, Harvey, Friedman, Hakel y Cornelius (1988) descubrieron que los factores del JEI tienen un estrecho paralelo con los del PAQ. El primero parece lograr las mismas metas que el segundo, pero con mayor velocidad y eficiencia.

En años recientes han aparecido varios cuestionarios estandarizados de análisis de puestos, como se analiza en el cuadro 11–2. En realidad, estos instrumentos aún están en su infancia. Sin embargo, con las reglamentaciones externas y la amenaza de las demandas por discriminación que siempre merodean en el ambiente, podemos predecir con confianza que los cuestionarios de análisis de puestos se volverán algo común —al menos dentro de grandes corporaciones que deben asignar diversos empleados a cargos específicos.

FUNCIÓN DE LAS PRUEBAS EN LA SELECCIÓN DE PERSONAL

COMPLEJIDADES DE LA SELECCIÓN DE PERSONAL

Con base en la suposición de que las pruebas y evaluaciones psicológicas puedan proporcionar

Cuadro 11–2. Cuestionarios de análisis de puestos

***Professional and Managerial Position Questionnaire (PMPQ*; Cuestionario de Puestos Profesionales y Gerenciales)**

(Mitchell y McCormick, 1979)

Un instrumento orientado a la gerencia que evalúa actividades interpersonales, planificación y toma de decisiones, procesamiento de información, actividades técnicas y factores antecedentes que constituyen requisitos previos como experiencia relacionada, entrenamiento especial y uso de un segundo idioma.

***Occupation Analysis Inventory (OAI*; Inventario de Análisis Ocupacional)**

(Cunningham, Boese, Neeb y Pass, 1983)

Un instrumento de 617 reactivos con propósitos generales que evalúa actividades biológicas/relacionadas con la salud, actividades botánicas, reparación eléctrica-electrónica, trabajo con animales y preparación/procesamiento de alimentos.

***General Work Inventory (GWI*; Inventario de Trabajo General)**

(Cunningham y Ballentine, 1982)

En esencia es una forma corta del Inventario de análisis ocupacional que evalúa constructos laborales similares.

***Common Metric Questionnaire (CMQ*; Cuestionario Métrico Común)**

(Harvey, 1990)

El CMQ se diseñó para aplicarse a todos los empleos; incluye escalas con referencia conductual, de modo que las calificaciones significan lo mismo entre diferentes puestos; por ejemplo, las interacciones con otras personas podrían calificarse en: coordinar/programar sus actividades, convencerlos o persuadirlos, entrenarlos/instruirlos/educarlos.

información valiosa sobre el desempeño potencial en el trabajo, muchos negocios, empresas y ambientes militares han utilizado las puntuaciones de prueba y los resultados de evaluación para la selección de personal. Como ha señalado Guion (1991), la investigación I/O sobre selección de personal ha enfatizado la validez relacionada con el criterio en contraposición con la validez de contenido o de constructo. Estos otros enfoques hacia la validez son, con toda seguridad, pertinentes, pero en general toman una posición subordinada con respecto a la validez relacionada con el criterio, que predica que los resultados actuales de la evaluación deben predecir el futuro criterio de desempeño en el cargo.

Desde el punto de vista de la validez relacionada con el criterio, la lógica de la selección de personal es atractivamente sencilla. Ya sea en grandes empresas o en pequeños negocios, aquellos que seleccionan a los empleados deberían utilizar pruebas o evaluaciones que tengan correlaciones documentadas y fuertes con el criterio del desempeño en el puesto y después contratar a individuos que obtengan las puntuaciones más altas en la prueba o que muestren los resultados más fuertes en la evaluación. ¿Qué podría ser más sencillo que esto?

Por desgracia, la aplicación en la vida real de los procedimientos de selección de personal está plagada de complejidades psicométricas y escollos legales. Las complicaciones psicométricas surgen, en gran medida, del hecho de que la conducta laboral es, en pocas ocasiones, una conducta sencilla y unidimensional. Existen algunas excepciones (como la producción en una línea de ensamblado), pero la regla general en nuestra sociedad posterior a la industrialización es que la conducta laboral es compleja y multidimensional. Incluso los empleos que parecen sencillos pueden ser sumamente complejos. Por ejemplo, considérese aquello que se requiere para el desempeño eficaz en la entrega de correo en EUA. El individuo que lleva el correo seis días por semana debe hacer más que simplemente colocarlo en el buzón. Esta persona debe clasificar de manera precisa y sobre la marcha el correo, interpretar e imponer los reglamentos del gobierno acerca del tamaño de los paquetes, manejar animales molestos e incluso peligrosos, reconocer y evitar peligros físicos y ejercer habilidades interpersonales eficaces para tratar con el público; por citar sólo unas cuantas de las complejidades de este puesto.

Por tanto, la selección de personal es una tarea confusa, condicional e incierta. Guion (1991) ha destacado la dificultad para predecir

la conducta compleja a partir de pruebas simples. En primer lugar, la conducta compleja está, en parte, en función de la situación. Esto significa que es posible que incluso un enfoque óptimo para la selección no sea válido para todos los candidatos. Es bastante claro que la selección de personal no es una simple cuestión de aplicar pruebas y consultar puntuaciones de corte.

Se debe reconocer también el profundo impacto de los edictos legales y reglamentarios sobre las prácticas de prueba I/O. Dado que dichas prácticas pueden tener consecuencias serias —por ejemplo, la determinación de a quién se contrata o se asciende— no causa sorpresa enterarse que los precedentes legales y los decretos de regulación restringen de manera estricta las prácticas de prueba I/O. En la sección final de este capítulo se revisan estas cuestiones.

ENFOQUES PARA LA SELECCIÓN DE PERSONAL

En reconocimiento de que la entrevista es una forma ampliamente utilizada de evaluación de personal, es adecuado concluir que la evaluación psicológica es casi una práctica universal en las decisiones de contratación. Incluso a partir de una definición estrecha que incluya sólo medidas de lápiz y papel, cuando menos dos terceras partes de las compañías en EUA participan en exámenes del personal (Schmitt y Robertson, 1990). Para propósitos de selección de personal, el psicólogo I/O puede recomendar uno o más de los siguientes elementos:

- Datos autobiográficos.
- Entrevista de empleo.
- Pruebas de capacidad cognitiva.
- Pruebas de personalidad, temperamento y motivación.
- Pruebas de integridad de lápiz y papel.
- Pruebas sensoriales, físicas y de destreza.
- Muestra de trabajo y pruebas situacionales.

Se volverá ahora la atención a un análisis breve de las pruebas y enfoques de evaluación típicos dentro de cada una de estas categorías. Se cierra el tema con una discusión acerca de las cuestiones legales en los exámenes de personal.

DATOS AUTOBIOGRÁFICOS

Según Owens (1976), las formas de solicitud de empleo que requieren de los antecedentes personales y laborales así como de datos demográficos como edad y estado civil, se han utilizado en la industria desde 1894. Por lo común, los datos autobiográficos objetivos o calificables —llamados en ocasiones **datos biográficos**— se obtienen por medio de un formato estructurado que de manera diversa se ha llamado formulario de información biográfica, forma de datos biográficos, formulario de solicitud, guía de entrevista, encuesta de antecedentes individuales o algún instrumento similar. Aunque es posible que el público lego no reconozca estos instrumentos como verdaderas pruebas con poder de predicción, los psicólogos I/O han sabido desde hace algún tiempo que los datos biográficos proporcionan una base excepcionalmente poderosa para la predicción del desempeño del empleado (Drakeley, Herriot y Jones, 1988; Cascio, 1976; Ghiselli, 1966; Hunter y Hunter, 1984; Reilly y Chao, 1982). Un hito importante en el enfoque de los datos biográficos es la reciente publicación del *Biodata Handbook*, un detallado estudio sobre la utilización de la información biográfica en la selección y en la predicción del desempeño (Stokes, Mumford y Owens, 1994).

La fundamentación para el enfoque de los datos biográficos es que puede predecirse la conducta futura relacionada con el trabajo a partir de las elecciones y logros pasados. Los datos biográficos tienen poder de predicción debido a que ciertos rasgos de carácter que son esenciales para el éxito también son estables y duraderos. Es probable que el joven consistentemente ambicioso que ha obtenido distinciones y logros en el bachillerato, continúe con este patrón hasta su adultez. Así, el solicitante

de empleo que ha fungido como editor de un periódico escolar en bachillerato —y que responde a un reactivo de datos biográficos sobre esta cuestión— con toda probabilidad es un mejor candidato para la gerencia de una empresa que aquel que no informa actividades extracurriculares en un formato de datos biográficos.

NATURALEZA DE LOS DATOS BIOGRÁFICOS

En general, los datos biográficos demandan datos "objetivos"; no obstante, en ocasiones se incluyen reactivos que detectan actitudes, sentimientos y juicios de valor. Excepto por datos demográficos como edad y estado civil, los reactivos de datos biográficos siempre se refieren a logros y acontecimientos pasados. Ejemplos de preguntas de este tipo incluyen el mayor nivel educativo obtenido, número de premios o reconocimientos especiales, ingresos en un empleo previo, número de empleos desempeñados en los últimos 10 años, extensión de tiempo en la presente dirección, número de publicaciones y de proyectos dirigidos en los últimos cinco años.

Una vez que se obtienen los datos biográficos, el psicólogo I/O debe diseñar un medio para predecir el desempeño laboral a partir de esta información. La estrategia más común es una forma de codificación empírica no muy diferente de la que se utiliza en las pruebas de personalidad. A partir de una muestra grande de trabajadores ya contratados, el psicólogo I/O designa a un grupo exitoso y a uno que no lo es con base en el desempeño, puesto, salario o calificaciones del supervisor. Entonces, se contrastan los datos biográficos individuales para estos dos grupos, a fin de determinar cuáles reactivos discriminan de manera más precisa entre trabajadores exitosos y los que no lo son. Los reactivos que tienen un fuerte poder de discriminación reciben ponderaciones elevadas en el esquema de puntuación. Por tanto, los nuevos solicitantes que responden a los reactivos en la dirección codificada reciben puntuaciones altas en el instrumento de datos biográficos y se predice que tendrán éxito. La validación cruzada del esquema de puntuación en una segunda muestra de trabajadores exitosos y no exitosos es un paso crucial para garantizar la validez del método de selección con datos biográficos. Los lectores que deseen tener mayores detalles sobre los métodos de codificación empírica para los instrumentos de datos biográficos deberán consultar a Murphy y Davidshofer (1988) y Stokes y colaboradores (1994).

VALIDEZ DE LOS DATOS BIOGRÁFICOS

Varios investigadores han examinado la validez de los datos biográficos y han llegado en general a conclusiones positivas (Mumford y Owens, 1987; Reilly y Chao, 1982; Rothstein, Schmidt, Erwin, Owens y Sparks, 1990; Russell, Mattson, Devlin y Atwater, 1990; Stokes y cols., 1994). Uno de los primeros estudios de Cascio (1976) es típico de estas conclusiones. Este investigador utilizó un instrumento muy sencillo de datos biográficos —una combinación ponderada de 10 reactivos de un formulario de solicitud de empleo— para predecir la rotación del personal secretarial femenino en una compañía de seguros de mediano tamaño. Las correlaciones por validación cruzada entre las calificaciones en datos biográficos y la duración en el puesto fueron de .58 para miembros de minorías y de .56 para personas no pertenecientes a ellas.* Drakeley y colaboradores (1988) hicieron una comparación entre datos biográficos y pruebas de capacidad cognitiva como medios de predicción del éxito del entrenamiento. Las calificaciones de datos biográfi-

* El lector interesado podría querer saber cuáles 10 reactivos de datos biográficos poseen tal poder de predicción. Éstos fueron edad, estado civil, edad de los hijos, educación, puesto en el empleo anterior, salario previo, amigo o familiar en la compañía, lugar de residencia, posesión de casa propia y extensión de tiempo en la presente dirección. Por desgracia, Cascio (1976) no revela las ponderaciones relativas o dirección de calificación para los reactivos.

cos tuvieron la misma validez predictiva que las pruebas cognitivas. Lo que es más, cuando se añadió a la ecuación de regresión la información de los datos biográficos mejoró la precisión predictiva de las pruebas cognitivas.

En un extenso estudio de investigación, Reilly y Chao (1982) compararon ocho procedimientos de selección en cuanto a validez y al impacto adverso sobre las minorías. Los procedimientos fueron datos biográficos, valoración de compañeros, entrevistas, autoevaluaciones, verificaciones de credenciales, logros académicos, juicios de expertos y técnicas proyectivas. Sin dejar de señalar que las pruebas de capacidad con estandarización adecuada proporcionan el procedimiento más equitativo y válido de selección, Reilly y Chao (1982) concluyeron que sólo los datos biográficos y las valoraciones de compañeros tenían validez sustancial igual a aquéllas de las pruebas estandarizadas. Por ejemplo, en la predicción de la productividad en ventas, el coeficiente promedio de validez de los datos biográficos fue un muy robusto .62.

Es necesario mencionar ciertas advertencias con respecto a los enfoques de datos biográficos en la selección de personal. Es posible que la ley le prohíba a los empleadores hacer preguntas sobre edad, raza, sexo, religión y otras cuestiones personales —incluso si se puede demostrar a nivel empírico que dichos datos predicen el desempeño en el puesto. También, aunque la incidencia de engaño es muy baja, no hay duda que los solicitantes astutos pueden adulterar los resultados en una dirección favorable. Como ocurre con cualquier instrumento de medición, los reactivos de datos biográficos requerirán una reestandarización periódica. Por último, un inconveniente potencial del enfoque de datos biográficos es que, por su misma naturaleza, este método captura el *status quo* de la organización y, por tanto, podría anular la innovación. Becker y Colquitt (1992) analizan las precauciones en el desarrollo de los formatos de datos biográficos.

Por tanto, existen pocas dudas acerca de que la información puramente objetiva de los datos biográficos puede predecir aspectos del desempeño laboral con bastante precisión. No obstante, es más probable que los empleadores dependan de la información subjetiva como las impresiones de entrevista cuando toman decisiones sobre contratación. A continuación se vuelve la atención hacia la investigación sobre la validez de la entrevista de trabajo en el proceso de selección.

ENTREVISTA DE TRABAJO

En general, la entrevista de trabajo es sólo una parte del proceso de evaluación, pero la mayoría de los administradores la consideran como el componente crucial y definitorio de la contratación (Miner y Miner, 1978). Landy (1985) informa que, por lo común, las compañías entrevistan ¡de 5 a 20 personas por cada individuo que contratan! Si se considera la importancia de la entrevista y su terrible costo para la industria y las profesiones, el lector no debería sorprenderse de saber que miles de estudios tratan sobre la confiabilidad y validez de la misma. En este libro sólo pueden destacarse unas cuantas tendencias; pueden encontrarse reseñas más detalladas en Conway, Jako y Goodman (1995), Landy (1985), y Schmitt y Robertson (1990).

Los primeros estudios sobre la confiabilidad de la entrevista fueron bastante aleccionadores. En diversos estudios y reseñas, la confiabilidad se evaluaba por lo común al correlacionar las valoraciones de diferentes entrevistadores que tuvieron acceso a los mismos candidatos para el empleo (Wagner, 1949; Mayfield, 1964; Ulrich y Trumbo, 1965). La confiabilidad intercalificadores de docenas de estos primeros estudios se colocó por lo común a la mitad del rango de .50, un nivel muy bajo como para proporcionar evaluaciones precisas de los candidatos al empleo. Esta investigación también reveló que los entrevistadores estaban propensos a una predisposición de halo y otras influencias distorsionantes en sus percepciones de los candidatos. La predisposición de halo —la cual se analiza en el siguiente tema— es la tendencia a clasificar a un candidato en un nivel alto o bajo en

todas las dimensiones debido a una impresión global.

Posteriormente, los investigadores descubrieron que la confiabilidad de la entrevista podía incrementarse a nivel sustancial si ésta la llevaba a cabo de manera conjunta un panel, en vez de un solo entrevistador (Landy, 1985). Además, las entrevistas estructuradas en las que cada entrevistador le hacían las mismas preguntas a cada entrevistado también resultaban mucho más confiables que aquéllas no estructuradas (Borman, Hanson y Hedge, 1997; Campion, Pursell y Brown, 1988). En estos estudios se encontraron confiabilidades en el rango de .70 y superiores.

La investigación sobre la validez de la entrevista ha seguido el mismo curso evolutivo que se señala para la confiabilidad: las primeras investigaciones que examinaron las entrevistas no estructuradas fueron bastante pesimistas, mientras que la investigación posterior con enfoques estructurados produjo conclusiones más prometedoras. En estos estudios, la validez de la entrevista se evaluó de manera típica al correlacionar los juicios de entrevista con alguna medida del desempeño en el puesto. Los estudios iniciales sobre la validez de la entrevista produjeron pésimos resultados casi de manera uniforme, con coeficientes típicos de validez que rondaban la parte media del .20 (Arvey y Campion, 1982).

Tomando en cuenta que las entrevistas rara vez se utilizan de manera aislada, los primeros investigadores también analizaron el incremento en la validez, que es el aumento potencial en ésta cuando la entrevista se utiliza junto con otra información. Estos estudios se fundamentaron en la suposición optimista de que la entrevista contribuiría de modo positivo a la valoración del candidato cuando se le utilizara junto con puntuaciones de pruebas objetivas y datos sobre antecedentes. Por desgracia, las conclusiones iniciales fueron contrarias a esta suposición casi por completo (Landy, 1985).

En algunos casos, los intentos por probar el incremento en la validez de la entrevista demostraron justo lo contrario, lo que podría llamarse decremento en la validez. Por ejemplo, Kelly y Fiske (1951) establecieron que la información de entrevista de hecho disminuía la validez de las valoraciones de estudiantes de posgrado. En este primer estudio clásico, la tarea consistía en predecir el desempeño académico de más de 500 estudiantes del posgrado en psicología. Se utilizaron diversas combinaciones de credenciales (una forma de datos biográficos), calificaciones en pruebas objetivas y entrevistas, como base para las predicciones clínicas del desempeño académico. En el cuadro 11–3 se presentan los coeficientes de validez. El lector observará que las credenciales por sí solas proporcionaron una base mucho mejor para la predicción, que éstas más una entrevista de una hora. Los mejores medios de predicción se basaron en las credenciales y las calificaciones en pruebas objetivas; de hecho, añadir una entrevista de dos horas a esta información disminuía la precisión de las predicciones. Estas conclusiones destacaron la superioridad de la predicción actuarial (basada en fórmulas obtenidas de modo empírico), por encima de la predicción clínica (basada en impresiones subjetivas). En el último capítulo de este libro se sigue este debate entre el enfoque actuarial y el clínico.

Los estudios recientes que utilizan entrevistas cuidadosamente estructuradas, incluyendo entrevistas situacionales, proporcionan una

Cuadro 11–3. Coeficientes de validez para las clasificaciones con base en diversas combinaciones de información

Base para la clasificación	Correlación con el desempeño académico
Credenciales por sí solas	.26
Credenciales y una hora de entrevista	.13
Credenciales y puntuaciones de prueba objetiva	.36
Credenciales, puntuaciones de prueba y dos horas de entrevista	.32

Fuente: Basado en datos de Kelly, E. L. y Fiske, D. W. (1951). *The prediction of performance in clinical psychology.* Ann Arbor: University of Michigan Press.

imagen más positiva de la validez de la entrevista (Borman, Hanson y Hedge, 1997; Maurer y Fay, 1988; Schmitt y Robertson, 1990). Cuando se corrigen las conclusiones en cuanto a restricción de rango y falta de confiabilidad de las calificaciones del desempeño en el puesto, el coeficiente medio de validez para las entrevistas estructuradas resulta ser un impresionante .63 (Wiesner y Cronshaw, 1988). En un reciente metaanálisis, Conway, Jako y Goodman (1995) concluyeron que el límite superior para el coeficiente de validez de las entrevistas estructuradas era .67, mientras que para las no estructuradas el coeficiente de validez era de sólo .34. Razones adicionales para preferir las entrevistas estructuradas incluyen sus posibilidades de defensa legal en caso de una demanda (Williamson, Campion, Malo y colaboradores, 1997) y, para gran sorpresa, su mínima predisposición entre diferentes grupos raciales de solicitantes (Huffcutt y Roth, 1998).

A fin de llegar a niveles aceptables de confiabilidad y validez, las entrevistas estructuradas deben diseñarse con cuidadoso esmero. Considérese el protocolo que utilizaron Motowidlo y colaboradores (1992) en su investigación sobre entrevistas estructuradas para los puestos gerenciales y de mercadotecnia en ocho compañías de telecomunicaciones. Su formato de entrevista se basó en cuidadoso análisis de incidentes críticos en mercadotecnia y gerencia. A los posibles empleados se les hizo un conjunto de preguntas estándar sobre la manera en que habían manejado situaciones pasadas similares a estos incidentes críticos. Los entrevistadores estaban entrenados para hacer preguntas indagatorias discrecionales acerca de los detalles de la manera en que los solicitantes habían manejado estas situaciones. A lo largo de la entrevista, los entrevistadores tomaron copiosas notas. Después, se calificó a los solicitantes en escalas vinculadas con ejemplos conductuales. Por último, se combinaron estas calificaciones para producir una puntuación total de entrevista que se utilizó en las decisiones de selección.

En resumen, bajo condiciones cuidadosamente diseñadas, la entrevista puede proporcionar una base confiable y válida para la selección de personal. Sin embargo, como señalan Schmitt y Robertson (1990), no siempre están disponibles las condiciones previas indispensables para la validez de la entrevista. El problema esencial es que es posible que cada entrevistador valore sólo a un pequeño número de solicitantes, de modo que la estandarización de sus calificaciones no siempre es realista. Aunque la entrevista tiene un potencial de ser válida como técnica de selección, en su aplicación común, no estructurada, es probable que exista razón sustancial para preocuparse.

¿Por qué se utilizan las entrevistas? Si la entrevista típica, no estructurada, es tan poco confiable e ineficaz como base para la valoración de un candidato a un empleo, ¿por qué los administradores continúan teniéndola en tan alta estima? En su revisión sobre la entrevista de trabajo, Arvey y Campion (1982) destacan varias razones para la persistencia de la entrevista, incluyendo consideraciones prácticas como la necesidad de convencer al candidato de las ventajas del empleo y razones sociales como la susceptibilidad de los entrevistadores a la ilusión de validez personal. Otros han enfatizado la importancia de la entrevista para evaluar una buena correspondencia entre el solicitante y la organización (Adams, Elacqua y Colarelli, 1994; Latham y Skarlicki, 1995).

Es difícil imaginar que alguna vez la mayoría de los empleadores pudieran eliminar por completo la entrevista del proceso de investigación y selección. Después de todo, la entrevista sí satisface la simple necesidad humana de conocer a las personas que podrían contratarse. No obstante, con base en 50 años de investigación, es evidente que, a menudo, los datos biográficos y las pruebas objetivas proporcionan un fundamento más poderoso para la valoración y selección del candidato que las entrevistas estructuradas.

PRUEBAS DE CAPACIDAD COGNITIVA

La capacidad cognitiva puede referirse ya sea a un constructo general semejante a la inteligen-

cia o a una variedad de constructos específicos como habilidades verbales, capacidad numérica, percepción espacial o velocidad perceptual (Kline, 1993). Las pruebas de capacidad cognitiva general y las medidas de habilidades cognitivas específicas tienen muchas aplicaciones en la selección, valoración e investigación de personal. Dichas pruebas son rápidas, de bajo costo y fáciles de interpretar. Un amplio conjunto de investigación empírica ofrece un apoyo modesto, pero fuerte, para la validez y equidad de las pruebas estandarizadas de capacidad en la selección de personal (Gottfredson, 1986). Con seguridad, parece claro que a menudo las pruebas de capacidad proporcionan una base excelente para la selección para un puesto, al menos según criterios objetivos como la capacidad para predecir el desempeño en un empleo. Por ejemplo, Hunter y Hunter (1984) llevaron a cabo un metaanálisis de la investigación sobre la predicción del desempeño en el puesto y concluyeron que, para empleos de primer ingreso, no existe ningún medio de predicción (excepto por la muestra de trabajo) que supere la validez de las pruebas de capacidad, las cuales mostraron un coeficiente medio de validez de .54.

Un debate continuo dentro de la psicología I/O es si la evaluación para el empleo se logra mejor con pruebas sumamente específicas de la capacidad o con medidas de capacidad cognitiva general. La mayor parte de la evidencia parece sustentar la conclusión de que un factor general de inteligencia (el llamado factor *g*) es, en términos generales, un mejor medio de predicción del éxito en el entrenamiento y en el trabajo que las calificaciones en medidas cognitivas específicas —aun cuando se utilicen de manera combinada varias medidas cognitivas específicas. Por supuesto, esta conclusión es contraria al sentido común y a la evidencia anecdótica. Por ejemplo, Kline (1993) ofrece la siguiente estampa:

> La cuestión es que los factores *g* son importantes, pero también lo son estos otros factores. Por ejemplo, una *g* elevada es necesaria para ser un buen ingeniero o un buen periodista. Sin embargo, la elevada capacidad especial también es requisito para el primero, mientras que es un factor que confiere pocas ventajas para un periodista. No obstante, para este último la elevada capacidad verbal es obviamente útil.

Curiosamente, la investigación empírica proporciona sólo un apoyo dispar para esta postura (Gottfredson, 1986; Larson y Wolfe, 1995; Ree, Earles y Teachout, 1994). Aunque el tema continúa en debate, la mayor parte de los estudios sustenta el papel primordial de *g* en la selección de personal (Borman y col., 1997). Quizá la razón por la que, en general, *g* funciona mejor que los factores cognitivos específicos en la predicción del desempeño en el puesto, es que la mayor parte de éstos son factorialmente complejos en cuanto a sus requisitos, a despecho de los estereotipos (Guion, 1991). Por ejemplo, el ingeniero exitoso debe explicar sus ideas a otras personas y, por tanto, necesita de capacidad verbal así como de habilidades espaciales. Dado que las medidas de capacidad cognitiva general detectan muchas habilidades cognitivas específicas, a menudo una prueba general predice el desempeño en trabajos complejos tan bien o mejor que las medidas de habilidades específicas.

Literalmente cientos de pruebas de capacidad cognitiva están disponibles para la selección de personal, de modo que no es factible analizar aquí el rango completo de instrumentos. En vez de ello, se destacarán tres pruebas representativas: una que mide capacidad cognitiva general, otra que guarda relación con la evaluación de capacidades mecánicas, y una tercera que se dirige a una faceta sumamente específica del trabajo de oficina. Los tres instrumentos elegidos para revisión —la *Wonderlic Personnel Evaluation* (Evaluación Wonderlic de Personal), la *Bennett Mechanical Comprehension Test* (Prueba Bennett de Comprensión Mecánica) y la *Minnesota Clerical Test* (prueba Minnesota de Trabajos de Oficina) —son sólo ejemplos de los miles de pruebas de capacidad cognitiva disponibles para la selección de personal. Con frecuencia las tres pruebas se utilizan en ambientes de negocios y, por tanto, merecen comentario específico. En el cuadro 11–4 se muestran pruebas representativas de capacidad cognitiva que se encuentran en la selección de personal. Algunas perspectivas clásicas sobre

Cuadro 11–4. Pruebas representativas de capacidad cognitiva que se utilizan en la selección de personal

Pruebas de capacidad general

Shipley Institute of Living Scale (Escala del Shipley Institute of Living)

Wonderlic Personnel Evaluation (Evaluación Wonderlic de Personal)

Wesman Personnel Classification Test (Prueba Wesman de clasificación de personal)

Personnel Tests for Industry (Pruebas de personal para la industria)

Baterías de prueba de aptitudes múltiples

General Aptitude Test Battery (Batería de pruebas de aptitud general)

Armed Services Vocational Aptitude Battery (Batería de aptitudes vocacionales de las fuerzas armadas)

Test de aptitudes diferenciales

Employee Aptitude Test (Prueba de aptitud del empleado)

Pruebas de aptitudes mecánicas

Bennett Mechanical Comprehension Test (Prueba Bennett de comprensión mecánica)

Minnesota Spatial Relations Test (Prueba Minnesota de relaciones espaciales)

Revised Minnesota Paper Form Board Test (Prueba revisada Minnesota de tablero con formas de papel)

SRA Mechanical Aptitudes (Aptitudes mecánicas SRA)

Pruebas de capacidad motora

Crawford Small Parts Dexterity Test (Prueba Crawford de destreza con pequeñas piezas)

Test no verbal de la Universidad de Purdue

Hand-Tool Dexterity Test (Prueba de destreza manual con herramientas)

Stromberg Dexterity Test (Prueba Stromberg de destreza)

Pruebas para trabajos de oficina

Minnesota Clerical Test (Prueba Minnesota de trabajos de oficina)

Clerical Abilities Battery (Batería de capacidades para trabajos de oficina)

General Clerical Test (Prueba general de trabajos de oficina)

SRA Clerical Aptitudes (Aptitudes para trabajos de oficina SRA)

Nota: SRA denota Science Research Associates. Estas pruebas se revisan en la serie *Mental Measurements Yearbook*.

el examen de la capacidad cognitiva pueden localizarse en Ghiselli (1966, 1973); Guion (1991), Hunter y Hunter (1984) y Reilly y Chao (1982). Borman y colaboradores (1997) y Murphy (1996) proporcionan un análisis actual sobre este tema.

PRUEBA WONDERLIC DE PERSONAL

Aunque se le describe como una prueba de personal, la *Wonderlic Personnel Test*, WPT (Prueba Wonderlic de Personal), es en realidad una prueba grupal de capacidad mental general (Hunter, 1989; Wonderlic, 1983). Lo que hace que este instrumento sea una especie de institución en la evaluación de personal es su formato (50 reactivos de opción múltiple), su brevedad (un límite de tiempo de 12 minutos) y sus numerosas formas paralelas (16 en la última cuenta). Los tipos de reactivos en la Wonderlic son bastante variados e incluyen vocabulario, reordenación de oraciones, solución de problemas aritméticos, inducción lógica e interpretación de proverbios. Los siguientes reactivos captan el sabor de la prueba Wonderlic:

1. REGRESIÓN es lo contrario de
 a. ingestión b. avance
 c. cierre d. apertura
2. Dos hombres compraron un automóvil que costó 550 dólares; *X* pagó 50 dólares más que *Y*. ¿Cuánto pagó X?
 a. $500 b.$300 c. $400 d. $275
3. RENDIR HENDIR —Estas palabras tienen
 a. significado similar b. Significado opuesto
 c. significado que no es ni igual ni opuesto

La confiabilidad de la WPT es impresionante, en especial si se considera la brevedad del instrumento. Las confiabilidades de consistencia interna llegan por lo común al .90, con confiabilidades de formas alternas que por lo general superan al .90. Se encuentran disponibles datos normativos con 126 000 adultos con edades de 20 a 65 años. En cuanto a la validez, si se considera que la WPT es una prueba breve de la capacidad mental general, las conclusiones

son bastante positivas (Dodrill y Warner, 1988). Por ejemplo, Dodrill (1981) informa de una correlación de .91 entre las calificaciones en la WPT y aquéllas en el WAIS. Esta correlación es tan elevada como aquélla encontrada entre cualesquiera dos pruebas establecidas de la inteligencia general. Hawkins, Faraone, Pepple, Seidman y Tsuang (1990) informan de una correlación similar ($r = .92$) entre la WPT y el CI del WAIS-R con 18 pacientes psiquiátricos crónicamente enfermos. No obstante, en su estudio, un individuo no pudo manejar el formato de la WPT, lo cual sugiere que el deterioro visoespacial grave puede invalidar la prueba. Una reciente innovación con la prueba Wonderlic es la adición de cuatro formas de la misma (denominadas *Scholastic Level Exam* [Examen para nivel escolar]) que se utiliza en la selección y orientación educativa. Todavía no se ha establecido de manera firme la validez de la prueba Wonderlic en ambientes educativos (Belcher, 1992).

Los revisores de la WPT expresan algunas preocupaciones acerca de las pautas interpretativas que aparecen en el manual de la prueba (Schoenfeldt, 1985). Por ejemplo, el manual sugiere que las personas que obtienen puntuaciones naturales entre 16 y 22 tienen una capacidad limitada para cualquier actividad que no sea una tarea rutinaria. Estas puntuaciones de la WPT corresponden a un CI de 93 a 104; es decir, dichas personas se encuentran dentro del rango normal de inteligencia por completo. Por tanto, las pautas de interpretación parecen tanto arbitrarias como innecesariamente restrictivas. El manual también lista las puntuaciones de corte que se utilizan en la industria para más de 75 ocupaciones, lo cual hace surgir el peligro de que un jefe de personal, sin el entrenamiento necesario, interprete de manera excedida las puntuaciones individuales. Esto sería especialmente problemático en el caso de minorías raciales, dado que las diferencias de este tipo en la WPT son sustanciales (Schoenfeldt, 1985).

Otra preocupación acerca de la prueba Wonderlic es que se penaliza injustamente a las personas evaluadas cuya lengua materna no es el inglés (Belcher, 1992). La prueba Wonderlic es con límite de tiempo. De hecho, depende en gran medida de la velocidad que se añaden puntos a los individuos de 30 años y mayores para compensar por la bien conocida disminución en velocidad que acompaña al envejecimiento normal. No obstante, no se ha hecho ninguna modificación para los angloparlantes no nativos que podrían tener también un desempeño más lento. Una solución a las diversas cuestiones de equidad que se citan antes consistiría en proporcionar normas para el desempeño sin límite de tiempo en la Wonderlic. Sin embargo, los editores se han resistido a esta sugerencia.

PRUEBA BENNETT DE COMPRENSIÓN MECÁNICA

En muchos empleos y ocupaciones, la comprensión de los principios mecánicos es un requisito previo para el desempeño exitoso. Los mecánicos de automotores, plomeros, ingenieros mecánicos, solicitantes de escuelas industriales y personas en muchas otras vocaciones "manuales" necesitan comprender los principios mecánicos básicos, a fin de tener éxito en sus campos. En estos casos, un instrumento útil para la valoración ocupacional es la *Bennett Mechanical Comprehension Test*, BMCT (Prueba Bennett de Comprensión Mecánica). Ésta consiste de láminas acerca de las cuales la persona evaluada debe responder preguntas directas. Las situaciones representadas enfatizan principios mecánicos básicos que podrían presentarse en la vida cotidiana. Por ejemplo, podría representarse una serie de bandas y volantes y se podría pedir al examinando que discerniera las revoluciones por minuto relativas de dos volantes. La prueba incluye dos formas equivalentes (S y T).

La BMCT se ha utilizado de manera amplia desde la segunda Guerra Mundial para la evaluación de personal militar y civil, de modo que existe un vasto cuerpo de datos técnicos y de validez para este instrumento. Los coeficientes de confiabilidad por mitades van desde .80 hasta la parte baja del rango de .90. Se proporcionan datos normativos extensos para varios

grupos. Con base en una cantidad enorme de información inicial, la validez concurrente y predictiva de la BMCT parece haberse establecido de manera adecuada (Wing, 1992). Por ejemplo, en un estudio con 175 empleados, la correlación entre la BMCT y la subprueba de Razonamiento Mecánico de la DAT fue un impresionante .80. Una conclusión fascinante es que la prueba resultó ser uno de los mejores medios de predicción del éxito de los pilotos durante la segunda Guerra Mundial (Ghiselli, 1966).

A pesar de su excelencia psicométrica, la BMCT necesita modernización. La prueba tiene una apariencia anticuada y muchos de los reactivos son obsoletos. Según estándares contemporáneos, algunos de los reactivos de la BMCT son sexistas o potencialmente ofensivos para las minorías (Wing, 1992). El problema con los reactivos obsoletos y ofensivos es que pueden sesgar sutilmente las calificaciones de la prueba. La modernización de la BMCT sería un proyecto sencillo que aumentaría la aceptabilidad de la prueba para mujeres y minorías, al mismo tiempo que preservaría su excelencia psicométrica.

PRUEBA MINNESOTA DE TRABAJOS DE OFICINA

La *Minnesota Clerical Test*, MCT (Prueba Minnesota de Trabajos de Oficina), la cual pretende medir la velocidad y precisión perceptual relacionadas con el trabajo de oficina, ha permanecido en esencia sin cambios en su formato desde su introducción en 1931, aunque las normas han pasado por varias revisiones, siendo la más reciente en 1979 (Andrew, Peterson y Longstaff, 1979). La MCT se divide en dos subpruebas: Comparación de números y Comparación de nombres. Cada subprueba consiste de 100 pares idénticos y 100 diferentes de combinaciones de dígitos o letras (cuadro 11–5). En general, los pares divergentes difieren sólo en un dígito o letra, de modo que la tarea de comparación representa un desafío. Se requiere que la persona examinada verifique sólo los pares idénticos, que están mezclados en forma aleatoria con pares diferentes. La puntuación depende de manera predominante de la velocidad, aunque se penaliza a la persona evaluada por los reactivos incorrectos (los errores se restan del número de reactivos correctos).

Cuadro 11–5. Reactivos similares a aquéllos encontrados en la Prueba Minnesota de trabajos de oficina

Comparación de números

1. 3496482 — 3495482
2. 17439903 — 17439903
3. 84023971 — 84023971
4. 910386294 — 910368294

Comparación de nombres

1. El Sol de México — El Sol de México
2. Semillas Velásquez — Semillas Velázquez
3. Juan G. Sánchez — Juan G Sánchez
4. Daniel González — Daniel González

La confiabilidad de la MCT es aceptable, con coeficientes informados de estabilidad en el rango de .81 a .87 (Andrew, Peterson y Longstaff, 1979). El manual también informa de gran cantidad de datos de validez, incluyendo algunas conclusiones que no son del todo halagadoras. En estos estudios, la MCT se correlacionó con medidas del desempeño en el puesto, de resultado de entrenamiento y calificaciones de pruebas relacionadas. El desempeño en el puesto de asistentes de director, empleados administrativos, mecanógrafos y cajeros de banco se correlacionó de manera significativa, pero no robusta, con las calificaciones en la MCT. Ésta también tiene altas correlaciones con otras pruebas de capacidad para trabajos de oficina.

No obstante, aún existen dudas acerca de la validez y rango de aplicación de la MCT. Ryan (1985) señala que el manual carece de un análisis acerca de los estudios importantes contra los que no lo son, relacionados con la validez. Además, los autores de la MCT no proporcionan información detallada acerca de los atributos

específicos de los empleos, pruebas o cursos que se utilizaron como medidas de criterio en los estudios de validez que se presentan. Por tal razón es difícil conjeturar exactamente qué es lo que mide la MCT. Tanto Thomas (1985) como Ryan (1985) se quejan de que las nuevas normas de 1979 son difíciles de utilizar debido a que los autores de la MCT proporcionan muy poca información acerca de la manera en que se constituyeron los diversos grupos normativos. Así, aunque el manual revisado de la MCT presenta nuevas normas para 10 categorías vocacionales, es posible que el usuario no esté seguro de cuál grupo normativo debe utilizar en su ambiente. Debido a las notables diferencias en desempeño entre los grupos normativos, la vaguedad de la definición representa un problema significativo para los usuarios potenciales de esta prueba.

PRUEBAS DE PERSONALIDAD Y TEMPERAMENTO

Varias pruebas de personalidad proporcionan una base útil para la selección de empleados, cuando se les utiliza de manera apropiada. Al citar prácticas pasadas, Muchinsky (1990) advierte en contra del empleo imprudente de las pruebas de personalidad para la selección de personal:

> Los inventarios de personalidad como el MMPI [Inventario Multifásico de la personalidad Minnesota] se utilizaron durante muchos años para la selección de personal —de hecho, se usaron mal o de manera excesiva. Se les empleó de manera indiscriminada para evaluar la personalidad de un candidato, aun cuando no se había establecido una relación entre las calificaciones de la prueba y el éxito en el puesto. Pronto, los inventarios de personalidad se vieron cuestionados.

Por supuesto, es esencial que las pruebas de personalidad posean un vínculo demostrado con el desempeño en el puesto, antes de que se les utilice para la selección de personal. Por desgracia, la literatura sobre este tema es un tanto aleccionadora para la mayor parte de las escalas de personalidad. Para ilustrar esta cuestión, en el cuadro 11–6 se han resumido las conclusiones clave de Hough, Eaton, Dunnette, Kamp y McCloy (1990). El cuadro presenta las correlaciones medias entre diversas categorías de medidas de personalidad y varios criterios relacionados con el puesto. Los datos se basan en miles de estudios publicados en el periodo de 1960 a 1984.

El lector verá que la mayor parte de las correlaciones son demasiado bajas como para tener valor práctico. Por ejemplo, la correlación media entre las puntuaciones de Adaptación (de diversas escalas de personalidad) y Destreza en el puesto (en general calificaciones del supervisor) es de sólo .13. Los datos para esta casilla se obtuvieron de 146 estudios diferentes, de modo que los resultados son sumamente estables e indican que, en promedio, las medidas de adaptación psicológica son medios muy débiles de predicción del desempeño laboral. No obstante, existen ciertas razones en estos estudios para tener un optimismo limitado, como lo indica la robusta correlación de –.43 entre Adaptación y Delincuencia (p. ej., descuido con los deberes laborales). Obsérvese también que las medidas de Fiabilidad se correlacionan en promedio en –.28 con Abuso de sustancias. En apariencia, es más fácil pronosticar ciertos criterios relacionados con el puesto que otros.

Se sabe que algunas pruebas tienen mayor validez que otras para aplicaciones específicas en la selección de personal. Por ejemplo, el Inventario Psicológico de California (CPI) proporciona una medida precisa del potencial para la gerencia (Gough, 1984, 1987). Ciertas escalas de la CPI predicen razonablemente bien el desempeño general de los estudiantes en academias militares (Blake, Potter y Sliwak, 1993). El *Inwald Personality Inventory* (Inventario de Personalidad Inwald) se ha validado de manera adecuada como prueba de investigación previa el empleo para las fuerzas policiacas (Inwald, 1988). El Inventario Multifásico de Personalidad Minnesota también merece mención como herramienta de selección para las fuerzas policiacas (Hiatt y Hargrave, 1988). Por último, el *Hogan Personality Inventory*, HPI (Inventario de Personalidad Hogan), se ha validado de modo adecuado para la predicción del desempeño laboral en ambientes militares, hospi-

Cuadro 11–6. Resumen de estudios de validez relacionada con el criterio con pruebas de personalidad para predecir criterios laborales

Constructo de personalidad	Criterios laborales			
	Participación en el puesto	Destreza en el trabajo	Delincuencia	Abuso de sustancias
Impetuosidad*	.04	.04	–.29	.06
Afiliación	.06	–.01	NA	–.03
Adaptación	.13	.13	–.43	–.07
Conformidad	.02	–.01	–.31	–.04
Fiabilidad	.17	.13	–.27	–.28
Reflexión†	–.10	.01	–.24	.18
Logro	.24	NA	–.35	NA
Masculinidad	.10	NA	.02	–.18
Locus de control	.25	NA	NA	NA

Nota: NA significa una ausencia de investigación sobre esta relación entre el medio de predicción y el criterio.
* Rasgo de personalidad que se caracteriza por sociabilidad, alegría y receptividad.
† Rasgo de personalidad que se caracteriza por análisis intelectual y una consideración detenida de las cosas.
Fuente: Adaptado con autorización de Hough, L. M., Eaton, N., Dunnette, M., Kamp, J. y McCloy, R. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities [Monografía]. *Journal of Applied Psychology, 75*, 581-595.

talarios y corporativos. El HPI se basó en la teoría de los Cinco Grandes factores de la personalidad (Tema 13A, Teorías y medición de la personalidad). Este instrumento se sometió a una validación cruzada relacionada con el criterio que dio por resultado una validez tan alta como .60 para algunas escalas (Hogan, 1986; Hogan y Hogan, 1986). En el Tema 14A, Inventarios de autoinforme, se analizan aplicaciones adicionales de las pruebas de personalidad en relación con el trabajo. Borman *et al.* (1997) proporcionan un buen resumen de los estudios recientes sobre este tema.

PRUEBAS DE INTEGRIDAD DE LÁPIZ Y PAPEL

Varios editores de pruebas han introducido instrumentos diseñados para detectar a los individuos propensos al robo y a otros candidatos indeseables como son las personas poco dignas de confianza o que se ausentan con frecuencia del trabajo (Hogan y Hogan, 1989; O'Bannon, Goldinger y Appleby, 1989; Ones y Viswesvaran, 1998; Ones, Viswesvaran y Schmidt, 1993; Sackett, Burris y Callahan, 1989). Estas pruebas se presentan en dos tipos claramente diferenciados: pruebas explícitas de integridad y medidas basadas en la personalidad. Se analizará cada tipo por separado, concentrando la atención en aquellas cuestiones que surgen de estas pruebas más que en detallar los méritos o deméritos de los instrumentos individuales. El cuadro 11–7 lista 20 de los instrumentos de empleo más común.

Un problema con las pruebas de integridad es que su naturaleza de instrumentos exclusivos dificulta examinarlas de la misma manera que a los tradicionales. En la mayor parte de los casos, las claves de calificación sólo están disponibles para los psicólogos de la institución que las utiliza, lo cual hace difícil la investigación independiente. No obstante, en la actualidad existe un cuerpo considerable de investigación sobre las pruebas de integridad, como se analiza en la siguiente sección sobre validez.

Cuadro 11–7. Pruebas de integridad de uso común

Pruebas explícitas de integridad

- *Accutrac Evaluation System* (Sistema de evaluación Accutrac)
- *Applicant Review* (Revisión del solicitante)
- *Compuscan*
- *Employee Attitude Inventory* (Inventario de actitudes del empleado)
- *Employee Reliability Inventory* (Inventario de confiabilidad del empleado)
- *Integrity Interview* (Entrevista de integridad)
- Orion Survey (Estudio Orion)
- *PEOPLE Survey* (Estudio PEOPLE)
- *Personnel Selection Inventory* (Inventario de selección de personal)
- *Phase II Profile* (Perfil fase II)
- *Reid Report and Reid Survey* (Informe Reid y Estudio Reid)
- *Rely*
- *Stanton Survey* (Estudio Stanton)
- *True Test* (Prueba de verdades)

Pruebas de integridad basadas en la personalidad

- *Employment Productivity Index* (Índice de productividad del empleo)
- *Hogan Personnel Selection Series* (Serie Hogan de selección de personal)
- Inwald Personality Inventory (Inventario Inwald de personalidad)
- *Personnel Decisions Inc., Employment Inventory* (Inventario de empleo de *Personnel Decisions Inc.*)
- *Personnel Outlook Inventory* (Inventario de actitud del personal)
- *Personnel Reaction Blank* (Formulario de reacción del personal)

Nota: Los editores y autores de estas pruebas pueden encontrarse en O'Bannon, R. M., Goldinger, L. A. y Appleby, G. S. (1989). *Honesty and integrity testing.* Atlanta, GA: Applied Information Resources.

PRUEBAS DE INTEGRIDAD

Las **pruebas de integridad** constan, por lo común, de dos secciones. La primera es una sección que trata sobre las actitudes hacia el robo y otras formas de deshonestidad, como las creencias acerca del alcance de los robos por parte de empleados, grado de condena hacia el robo, aprobación de las racionalizaciones comunes sobre el mismo y facilidad percibida de éste. La segunda es una sección que trata sobre las admisiones explícitas de robo y de otras actividades ilegales, como artículos robados en el último año, práctica de juegos de apuestas y uso de sustancias prohibidas. Las pruebas de este tipo que se han investigado más son el *Personnel Selection Inventory* (Inventario de selección de personal), el *Reid Report* (Informe Reid) y la *Stanton Survey* (Encuesta Stanton). El lector interesado puede encontrar las direcciones de los editores de estos y otros instrumentos relacionados en O'Bannon *et al.* (1989.)

En apariencia, las pruebas de integridad pueden falsearse de manera más sencilla que aquéllas basadas en la personalidad y, por tanto, podrían tener menor valor en la investigación de los solicitantes deshonestos. Por ejemplo, Ryan y Sackett (1987) crearon una prueba genérica de integridad, modelada a partir de los instrumentos existentes. La prueba contenía 52 reactivos sobre actitud y 11 sobre admisión. En comparación con un grupo de contrastación al que se le pidió que contestara con sinceridad y otro al que se le pidió que respondiera como solicitantes de empleo, los participantes a los que se les solicitó que "falsearan las respuestas para dar buena impresión" produjeron puntuaciones sustancialmente superiores (es decir, mejores actitudes y menos admisiones de robo).

PRUEBAS DE INTEGRIDAD BASADAS EN LA PERSONALIDAD

Por lo común, las pruebas de integridad basadas en la personalidad no contienen referencias obvias al robo o a otras formas de conducta indeseable de parte del empleado. Estas medidas son más sutiles en su enfoque y, por tanto, son menos ofensivas para la mayoría de los candidatos a un puesto. De hecho, algunas pruebas de integridad no son en realidad más que partes recicladas de pruebas existentes de la personalidad como el Inventario Psicológico de California (CPI). Por ejemplo, el *Personnel Reaction Blank* (Formulario de Reacciones del

Personal; Gough, 1971) se basa en aquellas partes del CPI que tratan sobre sociabilidad, fiabilidad, escrupulosidad, valores internos, continencia y aceptación de costumbres. En general, las medidas de lápiz y papel acerca de la escrupulosidad muestran fuertes relaciones con la integridad relacionada con el trabajo (Collins y Schmidt, 1993).

Una estrategia común de desarrollo de prueba para las medidas de integridad basadas en la personalidad es la codificación empírica contra un criterio de robo. El problema con este enfoque es el criterio: es raro que se hagan aprensiones por robo y las admisiones de éste pueden ser precisas o no. La tasa base de robo de empleados es casi imposible de establecer con precisión. Por ejemplo, las tasas de autoinforme de robo van de 28 a 62% en diferentes estudios (Camara y Schneider, 1994). Así, la clasificación de criterio de algunos individuos investigados puede no ser válida. Un segundo enfoque consiste en medir constructos amplios como la desviación general del empleado que se indica por la hostilidad hacia la autoridad, búsqueda de emociones, irresponsabilidad e insensibilidad social. Los instrumentos que emplean esta estrategia demuestran una capacidad moderada para predecir criterios globales como las calificaciones del supervisor acerca de la eficacia (Ones y cols., 1993; Sackett y cols., 1989).

Un problema grave con la mayor parte de las pruebas de integridad es la tasa elevada de error, que de manera común se encuentra en el rango de 30 a 60%. Debido a que, por lo general, las pruebas de integridad son la última barrera —se utilizan sólo con la pequeña fracción de solicitantes que tienen la capacidad necesaria y la experiencia pertinente— las organizaciones que emplean estas pruebas deben estar en posición de descartar a la mayoría de los solicitantes. Por supuesto, la elevada tasa de error es consecuencia, en parte, de las estrictas puntuaciones de corte, lo cual produce el rechazo de los empleados potencialmente valiosos (falsos positivos) junto con los verdaderos ladrones y truhanes (verdaderos positivos). De hecho, ésta es una cuestión de validez, como se analiza en la siguiente sección.

VALIDEZ DE LAS PRUEBAS DE INTEGRIDAD

Los editores de las pruebas de integridad han respondido a los psicólogos escépticos y al público desconfiado con una andanada de estudios de validez relacionada con el criterio. Ones y colaboradores (1993) le solicitaron datos sobre pruebas de integridad a las editoriales, autores y colegas. Estas fuentes resultaron ser sumamente cooperativas: los autores reunieron 665 coeficientes de validez basados en 25 pruebas de integridad aplicadas a más de medio millón de empleados. Con intrincados procedimientos de metaanálisis, Ones y colaboradores (1993) calcularon un coeficiente promedio de validez de .41 cuando las pruebas de integridad se utilizaron para predecir las calificaciones del supervisor acerca del desempeño laboral. Es interesante señalar que las pruebas de integridad predijeron las conductas de indisciplina en general (robo, actividades ilegales, ausentismo, retardos, abuso de sustancias prohibidas, despidos debidos a robo y violencia en el trabajo) mejor de lo que lo hicieron con el robo de parte de empleados por sí solo. Los autores concluyeron con un tenue respaldo para estos instrumentos:

> Cuando comenzamos con nuestra investigación sobre las pruebas de integridad, nos mostrábamos escépticos, al igual que muchos otros psicólogos industriales, acerca del uso de las pruebas de integridad en la industria. Ahora, con base en los análisis de una enorme base de datos consistente de más de 600 coeficientes de validez, concluimos que las pruebas de integridad tienen evidencia sustancial de validez generalizable.

Una serie de ingeniosos estudios de Cunningham, Wong y Barbee (1994) hacen eco de esta conclusión. Entre otros datos de apoyo, estos investigadores descubrieron que los resultados de pruebas de integridad se correlacionaban con la reintegración de excedentes en el pago de salarios —incluso cuando se había instruido a los individuos para que dieran una impresión positiva en la prueba de integridad.

Otros revisores se muestran más cautos en sus conclusiones. Al comentar las recientes re-

señas de la *American Psychological Association* y de la *Office of Technology Assessment*, Camara y Schneider (1994) llegaron a la determinación de que las pruebas de integridad no satisfacen las expectativas de los expertos en evaluación, pero que es probable que sean mejores que los métodos erráticos, sin estandarización, que utilizan muchos empleadores para elegir a los solicitantes.

Siguen existiendo varias preocupaciones acerca de las pruebas de integridad. Las editoriales pueden poner a la venta sus instrumentos para los usuarios no calificados, lo cual viola los estándares éticos de la *American Psychological Association*. Un segundo problema surge del desconocimiento de la tasa base de robo y de otras conductas indeseables, lo cual dificulta la identificación de puntuaciones óptimas de corte para las pruebas de integridad. Si éstas son muy estrictas, se descalificará de manera injusta a los candidatos honestos. Por el contrario, una puntuación muy indulgente de corte hace que la prueba sea inútil. Una última preocupación es que los factores de la situación pueden moderar la validez de estos instrumentos. Por ejemplo, la manera en que se presenta una prueba a las personas examinadas puede afectar poderosamente sus respuestas y, por tanto, sesgar la validez del instrumento.

Cada vez es más frecuente que las legislaturas y tribunales decidan el destino de las pruebas de selección a un empleo. Para los autores de pruebas y para el público, el riesgo es elevado: las compañías en EUA aplican un estimado de 5 millones de pruebas de integridad cada año. A partir de 1994, Massachusetts fue el único estado en prohibir las pruebas de integridad, pero cuando menos en otros seis la legislación está pendiente (Camara y Schneider, 1994). Es muy probable que la utilización de pruebas de integridad se restrinja cada vez más en los años por venir.

El debate sobre las pruebas de integridad enfrenta los intereses legítimos de las compañías contra los derechos individuales de los trabajadores. Con toda seguridad, las empresas tienen derecho a no contratar ladrones, drogadictos y agitadores. Pero en la búsqueda de esta meta, ¿cuál es el costo último para la sociedad de que se cuestione a millones de solicitantes de empleo acerca de sus conductas pasadas en cuanto a drogas, alcohol, conducta criminal y otras cuestiones sumamente personales? Hanson (1991) ha preguntado de manera retórica si se beneficia a la sociedad con el equilibrio actual de poderes —en el que las empresas pueden obtener información personal acerca de quién es supuestamente valioso y de quién no lo es. Parece casi inevitable que el Congreso de EUA entre en el debate. En 1988, el Presidente Reagan firmó el *Employee Polygraph Protection Act* (Decreto de protección de empleados contra el polígrafo) que eliminó efectivamente las pruebas de polígrafo en la industria (véase Tema 10B, Aplicaciones forenses de la evaluación). Quizás en los siguientes años se vea una reducción abrupta de las pruebas de integridad debido a un Decreto de protección de empleados contra las pruebas de integridad.

ESTRUCTURA Y MEDICIÓN DE HABILIDADES PSICOMOTRICES

Guion (1991) ha señalado la curiosa paradoja de que las capacidades sensoriales, físicas y de destreza sean esenciales para el éxito en muchas situaciones laborales y que, no obstante, éstas casi nunca se evalúen como parte de la investigación o selección de empleados. La falta de dichas pruebas no sólo es desafortunada desde los puntos de vista de la eficacia y equidad en la selección de empleados, sino que también hace surgir preocupaciones acerca de demandas legales, como se analizará más adelante. En esta sección se revisan de manera breve los instrumentos y cuestiones relacionadas con la evaluación de las habilidades psicomotrices. Pueden encontrarse perspectivas adicionales sobre este tema en Hogan y Quigley (1994) y en Blakley, Quinones, Crawford y Jago (1994).

Es importante una comprensión adecuada de los requisitos sensoriomotores de trabajos específicos por varias razones. Primero, los solicitantes con discapacidad física tienen una

base legítima para demandar legalmente si se les niega empleo debido a limitaciones psicomotrices que no tienen relación con el desempeño en el puesto. Por tanto, los empleadores están en una posición en la que necesitan documentar, por medio de un análisis de puestos, estudios de validez o ambos, que se requieren niveles mínimos de habilidad psicomotriz para el desempeño eficiente en el trabajo. En otro sentido, los empleados pueden demandar a la compañía si se lesionan o desarrollan problemas de salud en un puesto para el cual no poseen un nivel necesario de habilidad física. Como señala Guion (1991), las cuestiones sobre habilidad psicomotriz son demasiado importantes como para dejarlas al examen médico tradicional.

TAXONOMÍA DE LAS HABILIDADES PSICOMOTRICES

Fleishman y sus colaboradores han analizado intensamente durante varios años la medición de las habilidades psicomotrices (Fleishman, 1975; Fleishman y Quaintance, 1984). Con base en numerosos estudios analítico-factoriales, estos investigadores han identificado cerca de 20 factores específicos de habilidad en el dominio psicomotriz (Fleishman, 1975). Once de estos factores se denominan capacidades perceptomotrices, mientras nueve se llaman capacidades de destreza física (cuadro 11–8). Fleishman ha desarrollado escalas de calificación para ayudar a los empleadores a determinar las capacidades físicas requeridas para puestos específicos —el lector interesado puede consultar los detalles en Fleishman y Mumford (1988). En este caso se restringe la cobertura a pruebas más tradicionales que pueden utilizarse para la evaluación de las habilidades psicomotrices de empleados potenciales.

En algunas ocupaciones —en particular en oficios manuales— las pruebas de habilidades psicomotrices pueden utilizarse para selección o colocación. Sin embargo, los usuarios deberían demostrar la relevancia de los criterios de selección para el desempeño del puesto en sí

Cuadro 11–8. Taxonomía de habilidades psicomotrices

Capacidades perceptomotrices	Capacidades de destreza física
Precisión del control	Flexibilidad de extensión
Coordinación de múltiples miembros	Flexibilidad dinámica
Orientación de respuesta	Fortaleza estática
Tiempo de reacción	Fortaleza dinámica
Velocidad de movimiento de brazo	Fortaleza explosiva
Control de frecuencia (manejo de tiempo)	Fortaleza del tronco
Destreza manual	Coordinación corporal gruesa
Destreza de dedos	Equilibrio
Firmeza brazo-mano	Vigor
Velocidad muñeca-dedos	
Tino	

Fuente: Fleishman, E. A. (1975). Toward a taxonomy of human performance. *American Psychologist*, 30, 1127-1149.

(Hogan, 1991). A continuación se revisa una muestra de dichas pruebas.

ESTUDIO DE APTITUD DEL EMPLEADO

Dos de las 10 subpruebas del *Employee Aptitude Survey*; EAS (Estudio de Aptitud del Empleado) pueden utilizarse para evaluar habilidades perceptomotoras (Ruch y Ruch, 1980). El EAS es, principalmente, una batería cognitiva que se utiliza en la selección de personal dentro de organizaciones. No obstante, la Prueba 3 —Búsqueda Motora— y la Prueba 9 —Velocidad y Precisión Manual— captan varias de las habilidades perceptomotoras propuestas por Fleishman. La subprueba de Búsqueda Motora consta de 30 líneas entrelazadas con otras. La tarea consiste en trazar con rapidez cada línea desde el principio hasta el final, como en el diagrama de un circuito eléctrico. Cada línea termina en una de cinco letras —las opciones de respuesta. Se utilizan tanto las respuestas correctas (*C*) como

las incorrectas (*I*) para calcular la puntuación (*P*) con la fórmula $P = C - I/4$. La subprueba tiene un límite de tiempo de cinco minutos y parecería captar capacidades perceptomotoras importantes como precisión del control, destreza manual y firmeza de brazo-mano (cuadro 11–8). Esta subprueba se recomienda para seleccionar puestos como dibujante, ingeniero de diseño y técnico. La subprueba de Velocidad y Precisión Manual consiste en colocar un punto hecho con lápiz en tantas letras *O* como sea posible en un límite de tiempo de cinco minutos. Los puntos marcados deben estar dentro del círculo y los errores se sancionan. La fórmula para la puntuación corregida es $P = C - (5 \times I)$. Esta subprueba mide tino y otras capacidades perceptomotrices y se recomienda para la selección a puestos como el de trabajador administrativo, operario de maquinaria y trabajos que requieren precisión o tareas repetitivas.

El **Informe Técnico** (Ruch y Ruch, 1980) presenta buenas confiabilidades *test-retest* para las subpruebas (principalmente en el rango de .80) y proporciona datos normativos para 57 empleos o categorías de empleo específicos. También se proporcionan las calificaciones resumidas de más de 1 000 empleados varones de una compañía de manufacturas. La brevedad de las subpruebas de Búsqueda Motora y Velocidad y Precisión Manual (límite de tiempo de cinco minutos cada una) las hace opciones sumamente atractivas en la selección inicial para puestos técnicos y administrativos que requieren precisión en tareas rápidas. Para una revisión reciente del EAS véase Fitzpatrick (1984).

PRUEBA MINNESOTA DE TASA DE MANIPULACIÓN

La *Minnesota Rate of Manipulation Test* (Prueba Minnesota de Tasa de Manipulación) es un pilar venerable y respetado para la evaluación de la destreza dedos-mano-brazo. Esta prueba se ha utilizado desde el decenio de 1940 para la selección de empleados en una amplia variedad de ambientes industriales. La prueba consta de un tablero con 60 orificios con cilindros hechos a la medida, que son rojos de un lado y amarillos del otro. Las cinco subpruebas incluyen: Colocación, Vuelta, Desplazamiento, Vuelta y colocación con una mano, y Vuelta y colocación con ambas manos. Para cada subprueba se instruye a los examinandos a que inserten, volteen o muevan los cubos tubulares de maneras específicas con la mayor velocidad posible. Se proporcionan normas de referencia para una muestra de 11 000 adultos jóvenes y 3 000 adultos mayores. Esta prueba parecería medir las variables de Fleishman de velocidad de movimiento del brazo, destreza manual y destreza de dedos.

TEST NO VERBAL DE LA UNIVERSIDAD DE PURDUE

Otra prueba de habilidades psicomotrices que se utiliza ampliamente en la investigación previa al empleo es el Test no verbal de la Universidad de Purdue. Esta prueba se diseñó en la Universidad de Purdue en 1948 como auxiliar en la selección de empleados para diversos tipos de trabajo manual. La prueba mide destreza para dos tipos de actividad: movimientos gruesos de las manos, dedos y brazos; y destreza en la punta de los dedos necesaria para tareas de ensamblado. Esta prueba parecería evaluar una mezcla compleja de las habilidades psicomotrices que identificó Fleishman. En el cuadro 11–9 se describen varias pruebas adicionales de capacidades perceptomotrices.

Desde un punto de vista práctico, las pruebas psicomotrices como ésta pueden ser muy útiles para establecer los niveles mínimos de desempeño con propósitos de investigación previa al empleo. No obstante, su utilidad en la selección y colocación de personal es limitada. Uno de los principales problemas es que, por lo común, las pruebas psicomotrices muestran sustanciales efectos de práctica. Lo que es más, éstos son sumamente variables de un individuo a otro, lo que significa que, por lo común, la confiabilidad de estas pruebas es sólo modesta —las confiabilidades *test-retest* pocas veces

Cuadro 11–9. Pruebas de habilidades perceptomotrices

***Roeder Manipulative Aptitude Test* (Prueba Roeder de aptitud de manipulación):** clasificación y ensamblaje de tuercas, tornillos y arandelas.

***Pennsylvania Bi-Manual Worksample* (Muestra de trabajo bimanual de Pennsylvania):** ensamblaje de tuercas y tornillos.

***O'Connor Finger Dexterity Test* (Prueba O'Connor de Destreza de dedos):** colocación manual de pernos en orificios, como en una línea de ensamblado.

***O'Connor Tweezer Dexterity Test* (Prueba O'Connor de destreza con pinzas):** utilización de pinzas para colocar pernos individuales en orificios con diámetro de 1/16 de pulgada.

***Grooved Pegboard* (Tablero ranurado para clavijas):** rotación y colocación de clavijas en ranuras que tienen orientaciones aleatorias.

***Stromberg Dexterity Test* (Prueba Stromberg de destreza):** colocación de 54 discos circulares de colores (rojo, amarillo, azul) en una secuencia predeterminada.

Nota: Estas pruebas pueden obtenerse en la *Lafayette Instrument Company*, entre otras fuentes.

superan a .80 y a menudo son mucho más bajas, en los rangos de .50 y .60.

MUESTRA DE TRABAJO Y EJERCICIOS SITUACIONALES

Una **muestra de trabajo** es una réplica en miniatura del trabajo que han solicitado las personas evaluadas. Muchinsky (1990) señala que la meta del psicólogo I/O al diseñar una muestra de trabajo consiste en "tomar el contenido del puesto de la persona, reducirlo a un periodo manejable y permitir que los solicitantes demuestren su capacidad para realizarlo". Guion (1991) ha enfatizado que no es necesario que las muestras de trabajo incluyan todos y cada uno de los aspectos de un puesto, pero deberían enfocarse en los elementos más difíciles que discriminen de manera eficaz a los candidatos débiles de los fuertes. Por ejemplo, un puesto de oficinista-mecanógrafa puede incluir también el hacer café y mandados para el jefe. Sin embargo, éstas son tareas triviales que demandan tan poca habilidad que sería inútil incluirlas en una muestra de trabajo. Ésta debería examinar dominios importantes del puesto, no el universo completo del puesto.

Campion (1972) diseñó una ingeniosa muestra para mecánicos que ilustra el concepto anterior. Con las técnicas de análisis de puestos que se analizan al principio de este tema, Campion determinó que los aspectos esenciales para ser un buen mecánico se definían por el empleo exitoso de herramientas, precisión del trabajo y capacidad mecánica general. Con ayuda de mecánicos expertos, diseñó una muestra de trabajo que incorporaba estos aspectos del puesto a través de tareas típicas como instalar poleas y reparar una caja de velocidades. Se asignaron puntos a las conductas componentes para cada tarea. Los reactivos de ejemplo y sus ponderaciones correspondientes fueron los siguientes:

Instalación de poleas y bandas	**Calificación ponderada**
1. Antes de instalar, verifica la chaveta contra:	
___ Eje	2
___ Polea	2
___ Ninguno	0
Desarmado y reparación de caja de engranes	
10. Retira el cojinete viejo con:	
___ prensa y punzón	3
___ extractor de cojinetes	2
___ extractor de engranes	1
___ otro	0
Insertar un buje en la catalina y limar para ajustar a un eje	
4. Verifica el diámetro interno del buje contra el diámetro del eje:	
___ visualmente	1
___ calibrador de agujeros y micrómetros	3
___ calibrador Vernier	2
___ escala	1
___ no verifica	0

Campion encontró que el desempeño de 34 varones, mecánicos de mantenimiento, en la medida de muestra de trabajo se relacionaba de manera significativa y positiva con las valoraciones del supervisor acerca de su desempeño en el puesto, con coeficientes de validez de .42 a .66.

Un **ejercicio situacional** es aproximadamente el equivalente para empleados de una muestra de trabajo. Estos ejercicios se utilizan en gran medida para seleccionar a personas para puestos gerenciales y profesionales. La principal diferencia entre un ejercicio situacional y una muestra de trabajo es que el primero refleja sólo parte del puesto, mientras que el último es un microcosmos del trabajo completo (Muchinsky, 1990). En un ejercicio situacional se pide al posible empleado que se desempeñe bajo circunstancias que son sumamente similares al ambiente laboral anticipado. Se pueden obtener entonces medidas de logro como una base para estimar la probable productividad u otros aspectos de eficacia en el puesto. Los ejercicios situacionales con la mayor validez muestran una semejanza estrecha con el criterio; es decir, los mejores ejercicios son extremadamente realistas (Asher y Sciarrino, 1974; Muchinsky, 1990).

Las muestras de trabajo y los ejercicios situacionales se basan en la sabiduría convencional de que el mejor medio de predicción del desempeño futuro en un área específica es el rendimiento pasado en esa misma área. Por lo común, un ejercicio situacional requiere que el candidato se desempeñe en un ambiente que sea muy similar al ambiente laboral destinado. Así, las medidas de desempeño resultantes se asemejan a aquello que compone al trabajo eventual en sí.

A lo largo de los años se han desarrollado cientos de muestras de trabajo y ejercicios situacionales. Por ejemplo, en una reseña anterior, Asher y Sciarrino (1974) identificaron 60 procedimientos, incluyendo los siguientes:

- Prueba de mecanografía para personal de oficina.
- Prueba de ensamblaje mecánico para mecánicos de telares.
- Prueba de lectura de mapas para controladores aéreos.
- Prueba de destreza con herramientas para maquinistas y riveteadores.
- Prueba de organización de encabezados, diagramación e historias para editores de revista.
- Prueba oral de indagación para consultores en comunicación.
- Prueba de *role-playing* para personal de ventas telefónicas.
- Redacción de cartas de negocios para gerentes.

Un ejercicio situacional muy eficaz que se analizará aquí es la técnica de bandeja de entrada, un procedimiento que simula el ambiente de trabajo de un administrador.

PRUEBA DE BANDEJA DE ENTRADA

El artículo típico sobre la **Prueba de Bandeja de Entrada** es la monografía de Fredericksen (1962). Para este amplio estudio, Frederiksen diseñó la *Bureau of Business In-Basket Test* (Prueba de Bandeja de Entrada del Departamento Comercial), que consiste de cargas, memorandos, registros de llamadas telefónicas y otros documentos que se acumulan en la bandeja de entrada de un ejecutivo recién contratado de un departamento comercial. En esta prueba no se pide al candidato que represente un papel, sino que sea él mismo.* Éste no debe decir lo que haría, debe hacerlo.

Las cartas, memorandos, llamadas telefónicas y entrevistas que se realicen en este ambiente laboral simulado constituyen el registro de conducta que se califica tanto en referencia al contenido como al estilo de las respuestas. El **estilo de respuesta** se refiere a cómo se llevó a cabo la tarea —con cortesía, por teléfono, mediante la participación de un superior, dele-

* No intentamos promover aquí un sexismo sutil, pero de hecho Frederiksen (1962) evaluó a una muestra de estudiantes, administradores, ejecutivos y oficiales del ejército que de manera predominante (si no es que exclusiva) eran varones.

gando a un subordinado y demás. El **contenido** se refiere a lo que se hizo, incluyendo planes, establecimiento de fechas límite, búsqueda de información; también se calculan varios índices cuantitativos, incluyendo el número de reactivos intentados y el total de palabras escritas. Para algunos criterios de calificación, como capacidad de inventiva —el número de líneas de acción que parecieron buenas ideas— se requiere del juicio experto.

Frederiksen (1962) aplicó su prueba de bandeja de entrada a 335 individuos, incluyendo estudiantes, administradores, ejecutivos y oficiales del ejército. La calificación de la prueba fue un procedimiento complejo que requirió del desarrollo de un manual con 165 páginas. La confiabilidad pares-nones de los reactivos individuales varió considerablemente, pero surgieron suficientes reactivos con una confiabilidad modesta (*r* de .70 y superiores) como para que Frederiksen (1962) pudiera realizar varios análisis factoriales y también hacer comparaciones grupales significativas.

Cuando las calificaciones de los reactivos individuales se correlacionaron entre sí y después se sometieron a análisis factorial, se pudo describir la conducta de los administradores potenciales en términos de ocho factores primarios. Cuando las calificaciones de estos factores primarios en sí se sometieron a análisis factorial, surgieron tres factores de segundo orden. Éstos describen la conducta administrativa en los términos más generales posibles. La primera dimensión es Preparación para la acción, que se caracteriza por diferir las decisiones finales hasta que se obtenga información y consejo. La segunda es sencillamente Cantidad de trabajo, que representa las grandes diferencias individuales en el mero trabajo llevado a cabo. La tercera se denomina Búsqueda de orientación, donde las personas con altas puntuaciones parecen ser ansiosas e indecisas. Estas dimensiones se adecuan bien a la teoría existente acerca del desempeño del administrador y, por tanto, sustentan la validez de la tarea de Frederiksen.

Cuando éste comparó los grupos de individuos en las dimensiones calificables de la prueba de bandeja de entrada, surgieron varios atributos sobresalientes. Por ejemplo, los estudiantes de pregrado destacaron la productividad verbal, los administradores del gobierno carecían de preocupación por personas externas, los ejecutivos comerciales eran muy corteses, los oficiales del ejército exhibieron fuerte control sobre subordinados y los directores de escuela carecían de control firme. Estas diferencias grupales hablan en gran medida de la validez de constructo de la prueba de bandeja de entrada, ya que los resultados son consistentes con las expectativas teóricas acerca de estos grupos de participantes.

Los primeros estudios sustentan la validez predictiva de las pruebas de bandeja de entrada. Por ejemplo, Brass y Oldham (1976) demostraron que el desempeño en una prueba de bandeja de entrada tiene correspondencia con el desempeño en el puesto de los capataces, si se utilizan categorías apropiadas de calificación de la bandeja de entrada. En términos específicos, con base en esta prueba, los capataces que recompensan personalmente a los empleados por el trabajo bien hecho, que castigan por el trabajo deficiente, que establecen objetivos específicos de desempeño y que enriquecen los trabajos de sus subordinados, también reciben una calificación alta como administradores eficaces de parte de sus superiores. El poder de predicción de estas dimensiones de bandeja de entrada fue significativo, con coeficientes de correlación múltiple de .54 entre los medios de predicción y el criterio. En la actualidad pueden adquirirse pruebas estandarizadas de bandeja de entrada para utilizarlas en organizaciones privadas. Por desgracia, la mayor parte de estas pruebas son instrumentos internos de las empresas que no están disponibles para revisión general. A pesar de una reseña precautoria ocasional (p. ej., Brannick y cols., 1989), la técnica de bandeja de entrada aún se tiene en alta estima como método poderoso para valorar a candidatos a puestos gerenciales.

CENTROS DE EVALUACIÓN

Un centro de evaluación no es tanto un lugar como un proceso. Muchas empresas y departa-

mentos militares —así como unos cuantos gobiernos progresistas— han dedicado sitios especiales para la aplicación de técnicas de bandeja de entrada y ejercicios de simulación en el entrenamiento y selección de administradores. El propósito de un **centro de evaluación** consiste en valorar el potencial administrativo al exponer a los candidatos a múltiples técnicas de simulación que incluyen presentaciones en grupo, ejercicios de solución de problemas, ejercicios de discusión grupal, entrevistas y técnicas de bandeja de entrada. También se toman en consideración para la valoración general los resultados de pruebas tradicionales de aptitud y personalidad. Gerentes exitosos mayores y que tienen entrenamiento especial en técnicas de observación y valoración, observan y evalúan los diversos ejercicios de simulación. Los centros de evaluación se utilizan en una variedad de ambientes que incluyen negocios e industria, gobierno y el ejército. No hay duda de que un centro de evaluación con el diseño apropiado puede proporcionar una evaluación válida del potencial administrativo. La investigación de seguimiento ha demostrado que el desempeño de los candidatos en un centro de evaluación se correlaciona en gran medida con las calificaciones del supervisor para el desempeño en el trabajo (Gifford, 1991). Una pregunta que es más difícil responder es si los centros de evaluación son eficaces en cuanto a costos en comparación con los procedimientos tradicionales de selección. Después de todo, el financiamiento de un centro de evaluación es muy costoso. La pregunta clave es si el enfoque de centros de evaluación para la selección impulsa lo suficiente la productividad de la organización como para compensar por el gasto del proceso de selección. En un sentido anecdótico, la respuesta parecería ser un rotundo sí, dado que las decisiones inadecuadas de los administradores deficientes pueden ser muy costosas. No obstante, existen pocos datos empíricos que respondan a esta cuestión.

Goffin, Rothstein y Johnston (1966) compararon la validez de las pruebas tradicionales de personalidad (con la *Personality Research Form* [Forma de Investigación de la Personalidad], Jackson, 1984b) y del enfoque del centro de evaluación en la predicción del desempeño administrativo de 68 gerentes en una compañía de productos de silvicultura. Ambos métodos fueron equivalentes en la predicción del desempeño, lo cual sugeriría que el enfoque de centro de evaluación no compensa el costo adicional (muy sustancial). Sin embargo, cuando se utilizaron ambos métodos en combinación, la prueba de personalidad proporcionó un significativo incremento de validez por encima de aquél del centro de evaluación por sí solo. Así, las conclusiones de las pruebas de personalidad y de los centros de evaluación contribuyen cada una con información única que es útil en la predicción del desempeño. El ejemplo de caso 11–1 ilustra un centro de evaluación en acción.

RESUMEN

1. La psicología industrial y organizacional (psicología I/O) trata sobre la conducta en situaciones de trabajo (negocios, publicidad y el ejército). Los psicólogos I/O utilizan pruebas y evaluaciones psicológicas para propósitos diversos que incluyen contratación, colocación, promoción y valoración.
2. El análisis de puestos consiste en definir un puesto en términos de las conductas necesarias para llevarlo a cabo. Este análisis incluye dos componentes principales: descripción del puesto (características físicas y ambientales del trabajo) y especificación del mismo (características personales requeridas).
3. El *Position Analysis Questionnaire*, PAQ (Cuestionario de análisis de puestos), e instrumentos similares proporcionan información cuantificable pertinente para el análisis

EJEMPLO DE CASO 11–1
ASIGNACIÓN DE CARRERA EN CANADÁ

Gifford (1991) ha descrito ampliamente la manera en que el gobierno canadiense selecciona a sus administradores con un enfoque de centro de evaluación que se conoce como *Career Assignment Program* (CAP; Programa de asignación de carrera). Con base en entrevistas con ejecutivos eficaces, el equipo de investigación del CAP determinó 13 atributos vitales para el éxito ejecutivo:

Inteligencia	Planificación y organización
Creatividad	Capacidad para delegar
Tolerancia al estrés	Análisis y síntesis
Motivación	Juicio
Independencia eficaz	Comunicación oral
Liderazgo	Comunicación escrita
Relaciones interpersonales	

Para unos cuantos de estos atributos (p. ej., inteligencia, creatividad, comunicación escrita), las pruebas estandarizadas pueden proporcionar una base válida para la evaluación. No obstante, la mayor parte de estos atributos sólo pueden valorarse por medio de ejercicios de simulación gerencial. El equipo de investigación del CAP obtuvo como resultado un programa agotador de tres días que incluyó cinco ejercicios de simulación que se describen más adelante.

Al llegar al centro de evaluación, los candidatos reciben materiales por escrito que describen una entidad gubernamental hipotética. Cada candidato debe asumir el papel de Jefe ejecutivo de una unidad de reciente formación dentro de la entidad de gobierno. El proceso de evaluación consiste de una serie de cinco actividades y juntas con interacciones simuladas (con actores entrenados) que se necesitan para administrar la unidad dentro la entidad:

- **Tarea de bandeja de entrada** —respuestas a informes, mensajes telefónicos y memorandos dirigidos al ejecutivo.
- **Junta de personal** —el candidato se reúne con cuatro subordinados para resolver problemas que enfrenta la unidad y la entidad.
- **Junta informativa** —el candidato informa al presidente sobre el estado general de la unidad.
- **Presentación presupuestal** —presentación del presupuesto y previsiones a tres años, presentado a los principales funcionarios financieros.
- **Fuerza de trabajo** —el candidato ayuda a estudiar y recomendar soluciones para los problemas de la unidad y entidad.

El personal del CAP observa y valora todas estas actividades y entonces se califica a los candidatos en los 13 atributos antes listados. Estos datos se utilizan para determinar el potencial administrativo general. La investigación de seguimiento indica que el enfoque de centro de evaluación es válido a un nivel de predicción: las calificaciones del CAP se correlacionaron de manera elevada con las calificaciones del supervisor para el desempeño en el puesto hasta cuatro años después (Gifford, 1991).

de puestos. Por ejemplo, el PAQ evalúa los siguientes componentes de un puesto: obtención de información, procesos mentales, producto laboral, relaciones personales y contexto del puesto.

4. Las pruebas psicológicas pueden representar un papel importante en la selección de personal, pero se les debe utilizar con sensibilidad hacia las cuestiones de validez predictiva y preocupaciones legales. Los psicólogos I/O necesitan reconocer que es posible que incluso un enfoque óptimo de selección no sea válido para todos los candidatos.

5. Los datos autobiográficos, conocidos como datos biográficos, poseen validez predictiva sustancial para muchos tipos de selección de personal. En muchos estudios la validez predictiva de los datos biográficos (con valores en el rango de .50) rivaliza con la de las pruebas estandarizadas.

6. En la forma en que por lo común se utiliza en la selección de personal, la entrevista tiene baja confiabilidad y validez deficiente. Sólo cuando ésta se diseña en forma cuidadosa y tiene fuerte estructura puede proporcionar una base confiable y válida para la selección de personal.

7. Las pruebas de capacidad cognitiva proporcionan una fuerte base para la selección de personal en la mayor parte de las ocupaciones. Las pruebas de capacidad, cuyo único rival es la muestra de trabajo, tienen coeficientes de validez de .54 en promedio entre muchas pruebas y muestras.

8. Las pruebas cognitivas que miden capacidad general (*g*) a menudo predicen mejor el desempeño laboral que las medidas de capacidades específicas. La razón es que la mayor parte de los puestos son factorialmente complejos en cuanto a sus requisitos, lo cual garantiza que las medidas de *g* poseerán elevada validez predictiva.

9. Cuando se les valida para el uso propuesto, las pruebas de personalidad y temperamento pueden proporciona una base útil para la selección de personal. Por ejemplo, el *Hogan Personality Inventory,* HPI (Inventario de Personalidad Hogan), está bien validado para la predicción del desempeño laboral en ambientes militares, hospitalarios y corporativos.

10. Las pruebas de integridad de lápiz y papel se diseñaron para detectar a individuos propensos al robo y a otros candidatos indeseables. Algunos de estos instrumentos poseen validez predictiva moderada (p. ej., medidas basadas en la personalidad), pero su utilización hace surgir muchas dudas éticas.

11. La evaluación de las habilidades psicomotoras y sensoriales puede ser importante para algunas ocupaciones. Por ejemplo, la *Minnesota Rate of Manipulation Test* (Prueba Minnesota de tasa de manipulación) proporciona una medida de la destreza de dedos-mano-brazo que es útil para la investigación de empleados en una variedad de ambientes industriales.

12. Una muestra de trabajo es una réplica en miniatura del puesto que ha solicitado la persona evaluada. Aquélla diseñada adecuadamente (p. ej., a los posibles mecánicos se les podría pedir que instalen una polea y reparen una caja de velocidades) puede producir coeficientes de validez en los rangos de .40, .50 o .60.

13. Los ejercicios situacionales, como la prueba de bandeja de entrada, se utilizan principalmente para seleccionar a personas para los puestos gerenciales y profesionales. Aunque consumen mucho tiempo y son costosos, estos ejercicios proporcionan una base válida para la selección de gerentes.

14. Un centro de evaluación se utiliza para evaluar el potencial gerencial al exponer a los candidatos a múltiples técnicas de simulación, incluyendo presentaciones de grupo, ejercicios de solución de problemas, entrevistas y técnicas de bandeja de entrada. Las calificaciones del centro de evaluación ayudan a identificar el talento gerencial de alto nivel.

TÉRMINOS Y CONCEPTOS CLAVE

Análisis de puestos	p. 479	Ejercicio situacional	p. 500
Datos biográficos	p. 483	Prueba de Bandeja de Entrada	p. 500
Pruebas de integridad	p. 494	Centro de evaluación	p. 502
Muestra de trabajo	p. 499		

TEMA 11B
Valoración del desempeño en el trabajo

Funciones de la valoración del desempeño
Enfoques para la valoración del desempeño
Fuentes de error en la valoración del desempeño
Cuestiones legales en la evaluación I/O
Ejemplo de caso 11–2. Prácticas imprudentes de prueba en la investigación de empleados
Resumen
Términos y conceptos clave

La valoración del desempeño en el trabajo es fundamental para la operación exitosa de cualquier empresa u organización. En ausencia de realimentación significativa, los empleados no tienen idea de cómo mejorar; sin una evaluación útil, los administradores no tienen idea de cómo manejar al personal. Es difícil imaginar la manera en que una corporación, empresa u organización podría alcanzar una misión institucional, sin valorar el desempeño de sus empleados de una manera u otra.

Con frecuencia, los psicólogos industriales y organizacionales ayudan a diseñar escalas de calificación y otros instrumentos que se utilizan para la valoración del desempeño (Landy y Farr, 1983). Cuando se hace de manera apropiada, la valoración de los empleados descansa sobre una sólida fundamentación de medidas de psicología aplicada —de aquí que se incluya como uno de los principales temas en este texto. Además de dar una introducción a los aspectos esenciales en la medición del desempeño en el trabajo, también se tocan de manera breve las muchas cuestiones legales que rodean a la selección y estimación del personal. Se comienza con un análisis del contexto de la valoración del desempeño.

FUNCIONES DE LA VALORACIÓN DEL DESEMPEÑO

La valoración del desempeño en el trabajo satisface muchos propósitos institucionales. La lista corta incluye ascensos, transferencias, despidos y el establecimiento de salarios —todo lo cual puede estar supeditado con esta tarea. La lista extensa incluye los 20 usos comunes que identificaron Cleveland, Murphy y Williams (1989):

- Administración de salarios
- Ascensos
- Retención o terminación
- Reconocimiento del desempeño individual
- Despidos
- Identificación del desempeño deficiente
- Identificación de las necesidades individuales de entrenamiento
- Realimentación del desempeño
- Determinación de transferencias y nombramientos
- Identificación de fortalezas y debilidades individuales
- Planificación de personal
- Determinación de necesidades de entrenamiento de la organización
- Valoración del logro de metas
- Asistencia en la identificación de metas
- Valoración de sistemas de personal
- Fortalecimiento de la estructura de autoridad
- Identificación de necesidades de desarrollo de la organización
- Criterios para investigación de validez
- Documentación de decisiones sobre el personal
- Satisfacción de requisitos legales

Estas aplicaciones de la valoración del desempeño se agrupan alrededor de cuatro usos principales: comparación de individuos en términos de sus niveles generales de desempeño; identificación y utilización de información acerca de fortalezas y debilidades individuales; puesta en marcha y valoración de sistemas de recursos humanos en organizaciones; y documentación o justificación de decisiones sobre el personal. Más allá de cualquier duda, la valoración del desempeño es esencial para el mantenimiento de la efectividad de la organización.

Como pronto descubrirá el lector, la valoración del desempeño es un problema complejo para el que las soluciones simples y obvias son, en general, incorrectas. En parte, la tarea es difícil debido a que los criterios del desempeño eficaz pocas veces son tan explícitos como "cantidad de artículos vendidos en dólares" (p. ej., para un vendedor) o "porcentaje de estudiantes que aprueban un examen nacional" (p. ej., para un maestro). Sin importar qué tanto pudieran preferirse los métodos eficaces para la evaluación de la efectividad de los empleados, a menudo los enfoques basados en el juicio son la única opción práctica.

Los problemas que se enfrentan cuando se pone en marcha una valoración del desempeño se conocen generalmente, en conjunto, como problema del criterio —una designación que apareció por primera vez en el decenio de 1950 (p. ej., Flanagan, 1956; Landy y Farr, 1983). La frase **problema del criterio** intenta trasmitir las dificultades implicadas en la conceptuación y medición de constructos del desempeño que, con frecuencia, son complejos, confusos y multidimensionales. Para un análisis amplio sobre este tema, el lector debería consultar las extensas reseñas de Austin y Villanova (1992), y Campbell, Gasser y Oswald (1996). En la siguiente sección se tocan algunos aspectos del problema del criterio.

ENFOQUES PARA LA VALORACIÓN DEL DESEMPEÑO

Literalmente existen docenas de enfoques distintos al nivel conceptual para la valoración del desempeño en el trabajo. En la práctica, éstos se descomponen en cuatro clases de información: medidas del desempeño como los cálculos de productividad; datos sobre el personal como la tasa de ausentismo; calificaciones de compañeros y autoevaluaciones; y valoraciones del supervisor como escalas de calificación. Como

se analizará después, las escalas de calificación que responden los supervisores son, con mucho, el método preferido de valoración del desempeño. Pero primero deben mencionarse de manera breve otros enfoques.

MEDIDAS DEL DESEMPEÑO

Las medidas del desempeño incluyen índices aparentemente objetivos como el número de ladrillos que coloca un albañil, ganancias totales de un vendedor o porcentaje de estudiantes aprobados por un maestro. Aunque los cálculos de producción parecerían los métodos más objetivos y válidos para la medición del criterio, existen serios problemas con este enfoque (Guion, 1965), que incluyen:

- Es posible que la tasa de productividad no esté bajo el control del trabajador. Por ejemplo, el empleado en una empresa de comida rápida sólo puede vender lo que las personas ordenan y el trabajador de una línea de ensamblado sólo puede avanzar al mismo paso que sus compañeros.
- Los cálculos de producción no pueden aplicarse a la mayor parte de los empleos. Por ejemplo, no existen unidades relacionadas de producción para un profesor universitario, un juez o un recepcionista de un hotel.
- El énfasis en los cálculos de producción puede distorsionar la calidad del resultado. Por ejemplo, el farmacéutico en un emporio de medicamentos por correo puede surtir las recetas con los medicamentos incorrectos, si su trabajo se valora solamente en función de la productividad.

Otro problema es que los cálculos de producción pueden ser poco confiables, en especial a través de periodos cortos, y detectar sólo una pequeña proporción de los requisitos del empleo, incluso cuando parecen ser el criterio definitivo. Por ejemplo, el volumen de ventas podría parecer el criterio ideal para la mayor parte de los puestos de ventas; no obstante, un vendedor puede elevar sus ventas al mentir sobre las cualidades de los productos de la compañía. Las ventas pueden ser muy elevadas durante varios años —hasta que los consumidores insatisfechos demanden a la empresa. Con seguridad, la productividad es importante en este ejemplo, pero la corporación también debería querer que se evalúen factores interpersonales como la honestidad en las relaciones con el cliente.

DATOS SOBRE EL PERSONAL: AUSENTISMO

Los datos del personal como la tasa de ausentismo proporcionan otra base posible para la valoración del desempeño. Con toda seguridad, los empleadores tienen buenas razones para vigilar el ausentismo y reducirlo a través de incentivos apropiados. Steers y Rhodes (1978) calculan que el ausentismo cuesta cerca de ¡25 mil millones de dólares cada año en productividad perdida! No es sorprendente que el ausentismo sea una medida atractiva de criterio que se ha investigado de manera extensa (Harrison y Hulin, 1989).

Por desgracia, el ausentismo resulta ser una medida generalmente inútil del desempeño en el trabajo, excepto en casos extremos de ausencias injustificables flagrantes. Uno de los principales problemas es su definición. Landy y Farr (1983) listan 28 categorías de ausentismo, muchas de las cuales no tienen correlación con las demás. Diferentes tipos de ausentismo incluyen el programado contra el no programado, autorizado contra no autorizado, justificado contra injustificado, contractual contra no contractual, enfermedad contra no enfermedad, médico contra personal, voluntario contra involuntario, explicado contra no explicado, con compensación contra sin compensación, enfermedad certificada contra enfermedad eventual, ausencia de Lunes/Viernes contra ausencia a mitad de semana, e informado contra no informado. ¿Cuándo está realmente ausente del trabajo el empleado? Los criterios son muy evasivos.

Además, el ausentismo resulta ser una variable atrozmente poco confiable. Las correla-

ciones *test-retest* (tasas de trabajadores ausentes de dos periodos de extensión idéntica) son tan bajas como .20, lo cual significa que los empleados presentan tasas muy variables de un periodo al siguiente. Un problema relacionado con el ausentismo es que los trabajadores tienden a informarlo pocas veces en su propio caso y lo hacen en extremo en el caso de los demás (Harrison y Shaffer, 1994). Por último, para la mayoría de los trabajadores, las tasas son bastante bajas. En pocas palabras, el ausentismo es un método deficiente para evaluar el desempeño del trabajador, excepto para un pequeño porcentaje con ausencias injustificadas constantemente.

CALIFICACIONES DE COMPAÑEROS Y AUTOEVALUACIONES

Algunos investigadores han propuesto que las calificaciones de compañeros y las autoevaluaciones son sumamente válidas y constituyen un complemento importante para las calificaciones del supervisor. Gran cantidad de investigaciones tratan sobre esta cuestión, pero a menudo los resultados son confusos y contradictorios. No obstante, es posible listar varias generalizaciones (Harris y Shaubroeck, 1988; Smither, 1994).

- Los compañeros dan calificaciones más indulgentes que los supervisores.
- La correlación entre las autocalificaciones y las calificaciones del supervisor es mínima.
- La correlación entre las calificaciones de compañeros y las del supervisor es moderada.
- Los supervisores y subordinados tienen diferentes ideas sobre aquello que es importante en el trabajo.

En general, los estudiosos concluyen que las calificaciones de compañeros y las autoevaluaciones pueden tener una aplicación limitada con propósitos como el desarrollo personal, pero su validez no se ha establecido aún de manera adecuada como para justificar su utilización generalizada (Smither, 1994).

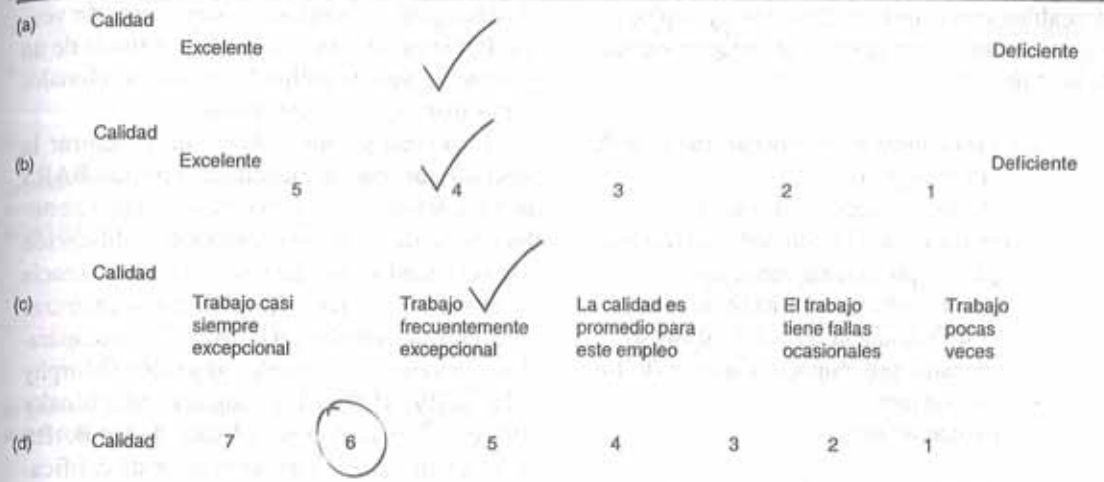
ESCALAS DE CALIFICACIÓN DEL SUPERVISOR

Las escalas de calificación son la medida más común del desempeño en el trabajo (Landy y Farr, 1983; Muchinsky, 1990). Estos instrumentos varían desde las simples formas gráficas hasta escalas complejas vinculadas con conductas concretas. En general, las escalas de calificación del supervisor revelan sólo una confiabilidad regular, con un coeficiente promedio de confiabilidad intercalificadores de .52 a través de muchos diferentes enfoques y estudios (Viswesvaran, Ones y Schmidt, 1996). A pesar de su débil confiabilidad, las calificaciones del supervisor todavía figuran como el enfoque de más uso. Cerca de tres cuartos de todas las valoraciones del desempeño dependen de métodos basados en el juicio como las escalas de calificación del supervisor (Landy, 1985).

La escala de calificación más simple es la escala de calificación gráfica, introducida por Donald Paterson en 1922 (Landy y Farr, 1983). Una **escala de calificación gráfica** consiste en etiquetas de rasgo, definiciones breves de dichas etiquetas y un continuo para la calificación. Como el lector observará en la figura 11–1, se han utilizado diversos tipos de escalas de calificación gráfica.

La popularidad de estas escalas se debe, en parte, a su sencillez; pero esto también es una de sus principales debilidades debido a que la dimensión de desempeño laboral que se valora puede haberse definido de manera vaga. La insatisfacción con las escalas de calificación gráfica condujo al desarrollo de muchos enfoques alternativos para la estimación del desempeño, como se analiza a continuación.

Una **lista de verificación de incidentes críticos** se basa en episodios reales de conducta deseable e indeseable en el trabajo (Flanagan, 1954). Por lo común, la persona que desarrolla una lista de verificación le pedirá a los empleados que construyan el instrumento al presentar ejemplos específicos de conducta laboral deseable e indeseable. Por ejemplo, supóngase que se tiene el propósito de desarrollar una lista de verificación para estimar el desempeño de los



Factores de calificación	Valoración del desempeño			
	Consistentemente excelente	Ocasionalmente excelente	Consistentemente promedio	Insatisfactorio
Calidad: Precisión Exactitud Claridad		✓		

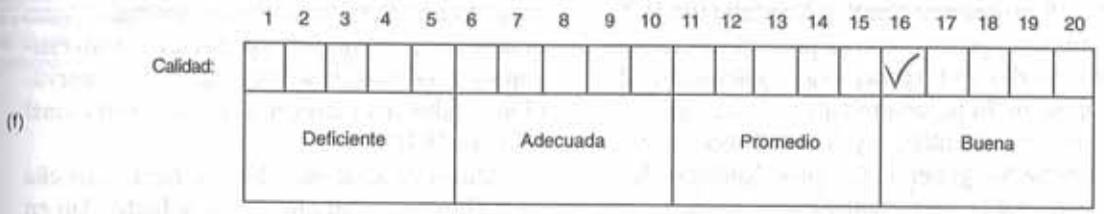


Figura 11–1. Ejemplos de escalas de calificación gráfica.

consejeros residentes (CR) en un dormitorio universitario. Si se sigue como modelo un estudio de Aamodt, Keller, Crawford y Kimbrough (1981), se les podría hacer la siguiente petición a los CR actuales del dormitorio:

> Considere al mejor CR que usted haya conocido alguna vez. Por favor describa en detalle varios incidentes que reflejen por qué esta persona fue el mejor consejero. Por favor haga lo mismo con el peor CR que haya conocido.

Con base en cientos de conductas nominadas, los autores de la lista de verificación deben proseguir con una condensación y codificación de estos incidentes para obtener un número más pequeño de conductas relacionadas, tanto

deseables como indeseables. Por ejemplo, los siguientes reactivos podrían ser adecuados para la lista de verificación de CR:

——permanece en el dormitorio más de lo requerido.
——rompe las reglas del dormitorio.
——es justo en relación con la disciplina.
——planea programas especiales.
——no disciplina a sus amigos.
——con frecuencia es poco amistoso.
——muestra preocupación acerca de los residentes.
——parece autoritario.

Por supuesto, la lista completa de verificación sería mucho más larga que ésta. El supervisor de los CR respondería a este instrumento como una base para la estimación del desempeño. De ser necesario, se puede extraer una calificación total de resumen a partir de ponderaciones apropiadas de los reactivos individuales. Harvey (1991) analiza las ventajas y desventajas de este enfoque.

Otra forma de medida de juicio referida al criterio es la *Behavirally Anchored Raitin Scale*; BARS (escala de **clasificación con base conductual**). El trabajo característico sobre las BARS procede de Smith y Kendall (1963). Estos autores propusieron un procedimiento complejo de desarrollo para producir juicios referidos al criterio. El procedimiento utiliza a varios expertos para identificar y definir dimensiones de desempeño, generar ejemplos conductuales y dar una escala significativa a las conductas. En general, el procedimiento es bastante complejo, consume mucho tiempo y es costoso. Se han sugerido numerosas variaciones y mejorías (Harvey, 1991). Una ventaja de las BARS y de otras escalas basadas en la conducta es su estricta sujeción a las pautas de la EEOC (*Equal Employment Opportunity Commission* [Comisión de Oportunidad Equitativa de Empleo]) que se analizará posteriormente en este capítulo. Las BARS y los enfoques relacionados se centran en las conductas más que en las características de personalidad o de actitud. En la figura 11–2 se presenta una escala con base conductual para el desempeño laboral de un supervisor de ventas. Por supuesto, la valoración profunda de un gerente de ventas incluiría escalas adicionales para otros aspectos del puesto.

La investigación sobre cómo mejorar la precisión de las calificaciones con las BARS ha dado resultados combinados. Algunos estudios encuentran menos errores de calificación —en especial una reducción en la indulgencia injustificada en las valoraciones— mientras otros no ven mejoría con las BARS en comparación con otros métodos de valoración (Murphy y Pardaffy, 1989). En general, Muchinsky (1990) concluye que el enfoque de las BARS no es mucho mejor que las escalas de calificación gráfica en la reducción de errores de calificación. No obstante, el proceso de desarrollo de escala de las BARS puede tener beneficios indirectos en cuanto a que obliga a los supervisores a prestar atención estrecha a los componentes conductuales del desempeño eficaz.

Una Behavior Observation Scale, BOS **(Escala de Observación Conductual)**, es una variación de la técnica de las BARS. La diferencia entre ambas es que el enfoque de la BOS utiliza un continuo desde "casi nunca" hasta "casi siempre", para medir la frecuencia con la que se desempeña un empleado en una tarea específica en cada dimensión conductual. Como ocurre con la técnica de las BARS, los investigadores cuestionan si las escalas de observación conductual merecen el esfuerzo adicional (Guion, 1991).

Una **escala de elección forzosa** se diseña para eliminar el prejuicio y la subjetividad en las calificaciones del supervisor al forzar una elección entre opciones que son iguales en cuanto a características socialmente deseables. En teoría, este enfoque hace imposible que el supervisor incline las calificaciones debido a prejuicio o subjetividad. Se utilizará la investigación líder de Sisson (1948) para ilustrar las características de este enfoque. Este autor desarrolló una escala para valorar a oficiales del ejército que consistía en tétradas de descripciones conductuales. Cada una contenía dos reactivos positivos y dos negativos equiparados en cuanto a cualidades socialmente deseables. Los cuatro

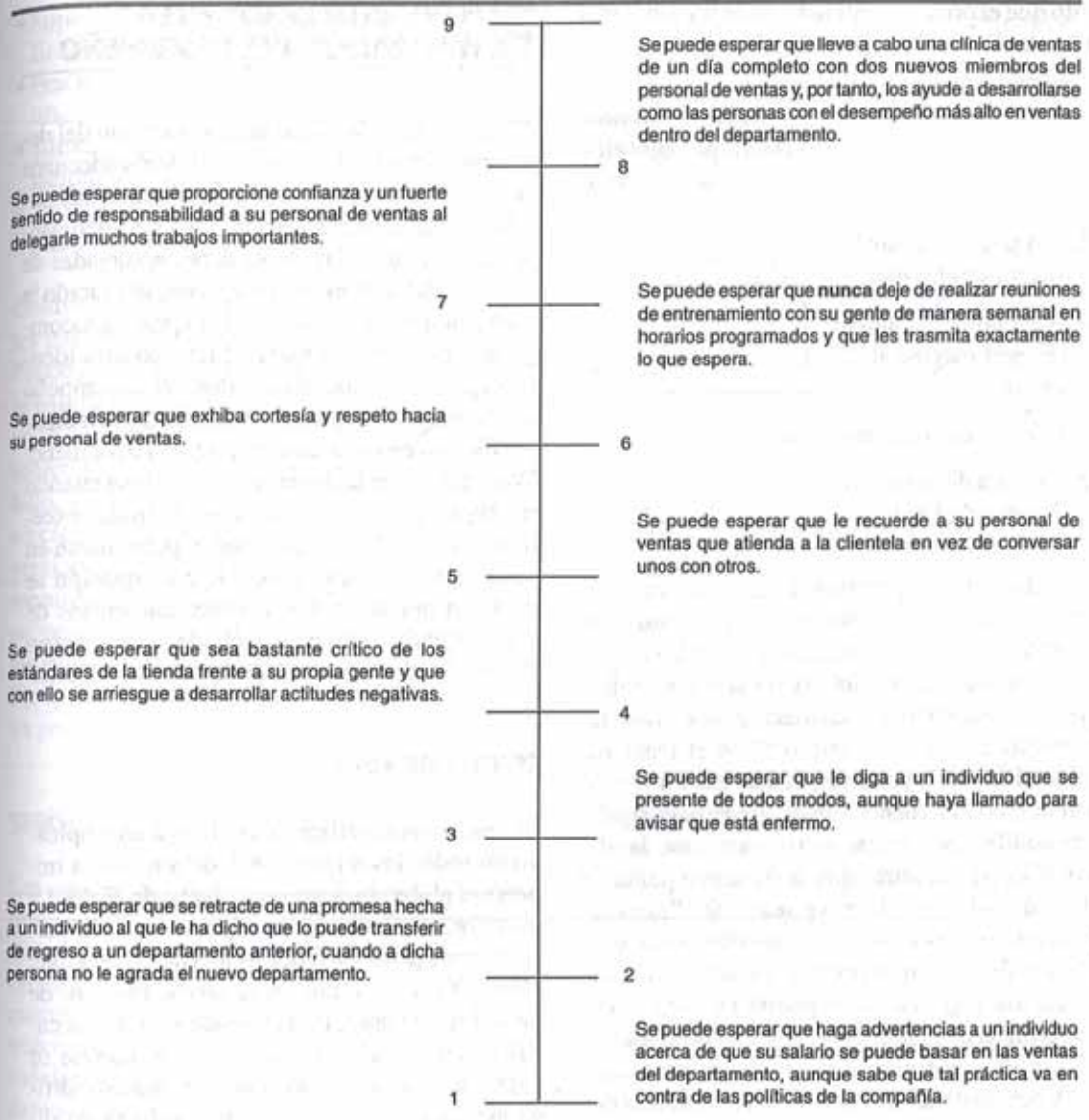


Figura 11–2. Escala de calificación con base conductual para supervisores de ventas.
Fuente: Reimpreso con autorización de Campbell, J. P., Dunnette, M.D., Arvey, R. D. y Hellervik, L. V. (1973). The development and evaluation of behaviorally based rating scales. *Journal of Applied Psychology, 57*, 15-22.

reactivos en cada tétrada se relacionaban temáticamente con una sola dimensión de desempeño. Sin que lo supieran los supervisores que respondían a la escala de calificación, uno de los dos elementos positivos se consideraba como muy descriptivo de los oficiales eficientes del ejército y el otro se juzgaba como menos descriptivo. De la misma manera, uno de los dos reactivos negativos se juzgaba como más descriptivo de los oficiales ineficientes del ejér-

cito que el otro. A continuación se presenta una tétrada de muestra (Borman, 1991):

	Más descriptivo	Menos descriptivo
A. No puede asumir responsabilidades	_____	_____
B. Sabe cómo y cuándo delegar responsabilidades	_____	_____
C. Ofrece sugerencias	_____	_____
D. Cambia de idea con demasiada facilidad	_____	_____

Se pidió a los supervisores que revisaran los reactivos en cada tétrada y que marcaran uno de ellos como el más descriptivo y otro como el menos descriptivo del oficial al que se valoraba. Se concedía una puntuación de +1 para la respuesta de "más descriptivo" en el reactivo con codificación positiva (en este caso la alternativa B) o de "menos descriptivo" para aquél con codificación negativa (en este caso la alternativa A), mientras que se daba una puntuación de –1 para la respuesta de "menos descriptivo" en el reactivo con codificación positiva o de "más descriptivo" en aquél con codificación negativa. La respuesta a los reactivos no codificados (alternativas C y D) como más o menos descriptivos obtenía una puntuación de 0. Así, cada tétrada producía un continuo de puntuaciones de cinco puntos: +2, +1, 0, –1, –2. La puntuación total que se utilizaba para la valoración del desempeño consistía en la suma algebraica de los reactivos individuales.

En realidad, el enfoque de elección forzosa nunca ha tenido gran éxito, debido en gran medida al esfuerzo que se requiere para la construcción de la escala. Esto es desafortunado, ya que el método sí reduce de manera eficaz el prejuicio injustificado. Borman (1991) considera este enfoque una "alternativa osada" que produce una escala de calificación relativamente objetiva.

FUENTES DE ERROR EN LA VALORACIÓN DEL DESEMPEÑO

El problema más difícil en la valoración del desempeño en el trabajo es la definición adecuada de los criterios de estimación. Si un supervisor utiliza un instrumento diseñado de manera deficiente que no detecta las dimensiones apropiadas de conducta laboral, entonces casi por definición la valoración del desempeño será imprecisa, incompleta y errónea. Sin duda, el fracaso para identificar los criterios apropiados del desempeño aceptable e inaceptable es una de las principales fuentes de error en la valoración del mismo. Pero ésta no es la única fuente. Incluso cuando los supervisores tienen acceso a medidas excelentes y bien diseñadas, pueden presentarse en ellas varios errores sutiles. A continuación se analizan tres de dichas fuentes adicionales de error de calificación: el efecto de halo, el sesgo del calificador y la contaminación del criterio.

EFECTO DE HALO

La tendencia a calificar alto o bajo a un empleado en todas las dimensiones debido a una impresión global se denomina **efecto de Halo**. La investigación sobre este efecto se puede rastrear hasta principios del siglo XIX (Thorndike, 1920). El más común es el efecto positivo de halo. En este caso, un empleado recibe una calificación superior a la que merece porque su supervisor no puede ser objetivo cuando califica aspectos específicos de la conducta de dicho empleado. En general, un efecto positivo de halo se debe a la generalización excesiva de un elemento de la conducta del trabajador. Por ejemplo, un empleado con asistencia perfecta puede recibir valoraciones mayores a las que merece en productividad y calidad del trabajo —aunque la asistencia no se relaciona de manera directa con estas dimensiones laborales.

Smither (1994) lista los siguientes medios para controlar los efectos de halo:

- Proporcionar entrenamiento especial a los calificadores.

- Supervisar a los supervisores durante la calificación.
- Practicar simulacros antes de realizar las calificaciones.
- Llevar un diario sobre la información relacionada con la estimación.
- Proporcionar a los supervisores una breve plática sobre los efectos de halo.

Goldstein (1991) examina enfoques adicionales para el entrenamiento de los calificadores. Puede encontrarse un fascinante análisis sobre la naturaleza y consecuencias del error de halo en Murphy, Jako y Anhalt (1993). Contrario al prejuicio reinante en contra de los errores de halo, estos investigadores concluyen que el efecto de halo no necesariamente disminuye la precisión de las calificaciones, y señalan que, a menudo, un supuesto efecto de halo es el subproducto de una verdadera superposición en las dimensiones que se califican. No es probable que el debate sobre dicho efecto se resuelva pronto (Arvey y Murphy, 1998).

SESGO DEL CALIFICADOR

Las fuentes potenciales de **sesgo del calificador** son tan numerosas que aquí sólo pueden mencionarse algunos ejemplos destacados. Los errores de indulgencia o severidad ocurren cuando un supervisor tiende a calificar a los trabajadores en los extremos de la escala. La indulgencia también es el producto de factores externos como el atractivo de un empleado. Los errores de severidad se refieren a la práctica de calificar todos los aspectos del desempeño como deficientes. En contraste, los errores de tendencia central se presentan cuando el supervisor califica a todos como cercanos al promedio en todas las dimensiones de desempeño. Los errores contextuales ocurren cuando el calificador valora a un empleado en el contexto de otro, en vez de basarse en el desempeño objetivo. Por ejemplo, la presencia de un vendedor obsesionado con su trabajo, el cual tiene un volumen de ventas extremadamente alto, podría causar que el supervisor de ventas califique al resto del personal de ventas con menos de lo que se merece.

Recientemente, los investigadores han prestado considerable atención a los posibles efectos sesgados del agrado o desagrado que siente un supervisor hacia un subordinado. Es sorprendente que la tendencia de las conclusiones es que el afecto del supervisor (agrado o desagrado) hacia empleados específicos no introduce sesgo del calificador. En general, el afecto fuerte en cualquier dirección representa información válida sobre un empleado. Así, es frecuente que las calificaciones de afecto se correlacionen en gran medida con las del desempeño, pero esto se debe a que ambos son consecuencia de qué tan bien o qué tan mal se desempeña el empleado en el puesto (Ferris, Judge, Rowland y Fitzgibbons, 1994; Varma, DeNisi y Peters, 1996). Otras formas de sesgo del calificador se analizan en Goldstein (1991) y en Smither (1994).

CONTAMINACIÓN DEL CRITERIO

Se dice que la **contaminación del criterio** existe cuando una medida de criterio incluye factores que no son parte demostrable del trabajo (Borman, 1991; Harvey, 1991). Por ejemplo, si una medida del desempeño incluye la apariencia, éste muy probablemente sería un caso de contaminación del criterio —a menos que la apariencia fuera relevante para el éxito en el trabajo. De la misma manera, valorar a un empleado en cuanto a "trato con el público" sería apropiado sólo si el puesto requiere en realidad de que éste trate con el público. Goldstein (1992) destaca tres tipos de contaminación del criterio:

1. El sesgo de oportunidad ocurre cuando los trabajadores tienen diferentes oportunidades de éxito, como cuando se asigna a un vendedor a un vecindario de altos ingresos y otros deben buscar sus ventas en áreas rurales aisladas.
2. El sesgo de características grupales está presente cuando éstas afectan al desempeño individual, como cuando los trabajadores en la misma unidad llegan a un acuerdo de li-

mitar su productividad para conservar relaciones sociales positivas.

3. El sesgo de conocimiento del medio de predicción ocurre cuando un supervisor permite que sus conocimientos personales sobre el empleado inclinen la estimación, como cuando la calidad de la universidad a la que asistió una nueva empleada afecta la evaluación que se hace de ella.

La atención cuidadosa en el análisis de puestos como base para la selección de los criterios de valoración es la mejor manera de reducir los errores en la valoración del desempeño. Además, los empleadores deben seguir ciertas pautas, como se analiza enseguida.

PAUTAS PARA LA VALORACIÓN DEL DESEMPEÑO

La valoración del desempeño es una tarea colosal. Los empleadores no sólo deben prestar atención a la solidez psicométrica de su enfoque, sino también deben diseñar un sistema práctico que satisfaga las metas de la organización. Por ejemplo, las normas de la valoración deben ser suficientemente difíciles y detalladas como para garantizar que se logren las metas de la organización. Otra preocupación es que la valoración del desempeño se encuentre dentro de los límites del Título VII del Decreto de 1964 sobre derechos civiles. Por tanto, los empleadores deben desarrollar sistemas equitativos que no discriminen debido a la raza, sexo y otras categorías protegidas. Para complicar las cosas, estas normas —solidez, cuestiones prácticas y legalidad— pueden entrar en conflicto entre sí. Es posible que el enfoque práctico no sea ni sólido en términos psicométricos ni legales. A menudo, los métodos de valoración que muestran las mejores características de medición (p. ej., fuerte confiabilidad intercalificadores) no evaluarán los aspectos más importantes del desempeño, es decir, no son prácticos. Éste es un tema familiar dentro del campo de la medición. Con demasiada frecuencia, los psicólogos deben optar entre rigurosidad y pertinencia, y en pocas ocasiones logran obtenerlas al mismo tiempo. Por último, deben tomarse en cuenta las consideraciones legales cuando se exploran los límites de la valoración del desempeño.

Smither (1994) ha publicado pautas para el desarrollo de sistemas de valoración del desempeño que se parafrasean aquí:

- Base la valoración del desempeño en un análisis cuidadoso de puestos.
- Desarrolle criterios de valoración específicos, libres de contaminación, a partir del análisis de puestos.
- Determine que el instrumento utilizado para calificar el desempeño sea apropiado para la situación de valoración.
- Entrene a los calificadores para que sean precisos, justos y legales en su utilización del instrumento de valoración.
- Emplee valoraciones del desempeño a intervalos regulares de seis meses a un año.
- Valore de manera periódica el sistema de valoración del desempeño para determinar si de hecho logra mejorar el desempeño.

El entrenamiento de los calificadores es una pauta de especial importancia. Un sistema de valoración que le parece perfectamente directo a un empleador, a un calificador no entrenado puede causarle confusión, lo cual resulta en un sesgo en las valoraciones. Borman (1991) señala que dos tipos de entrenamiento del calificador son eficaces: el entrenamiento en errores, en el que el entrenador simplemente busca alertar a los calificadores acerca de tipos específicos de error (p. ej., efecto de halo); y el entrenamiento de marco de referencia, en el que el entrenador familiariza a los calificadores con el contenido específico de cada dimensión de desempeño. La investigación indica que estos tipos de entrenamiento mejoran la precisión de las calificaciones.

CUESTIONES LEGALES EN EVALUACIÓN I/O

Casi todos los aspectos de la relación de trabajo están sujetos a la ley: reclutamiento, investi-

gación, selección, colocación, compensación, promoción y valoración del desempeño caen dentro del dominio de las interpretaciones legales (Cascio, 1987). No obstante, los tribunales y cuerpos legislativos han reservado un escrutinio especial para las pruebas relacionadas con el empleo. El profesional que se niega a conocer las pautas legales pertinentes en cuanto a pruebas de personal lo hace bajo un gran riesgo, ya que las prácticas imprudentes pueden conducir a litigios costosos y largos (ejemplo de caso 11–2).

Las pruebas de personal son particularmente sensibles debido a que las consecuencias de una decisión adversa son, a menudo, graves: El solicitante no obtiene el empleo o un empleado no logra un ascenso o colocación deseados. En reconocimiento a que las pruebas de empleo tienen una función sensible como guardianes que protegen el beneficio económico, el Congreso de EUA ha aprobado leyes que regulan de manera estricta el empleo de las pruebas. Los tribunales también han emitido decisiones que ayudan a definir lo que es una discriminación parcial de prueba. Además, los cuerpos reguladores han publicado pautas que tienen un impacto sustancial sobre las prácticas de prueba. En este tema se proporcionará una perspectiva actual sobre la regulación de las pruebas para el personal al seguir el desarrollo de las leyes, reglamentos y principales casos judiciales.

Ejemplo de caso 11–2
Prácticas imprudentes de prueba en la investigación de empleados

Según la información de Prensa Asociada del 11 de julio de 1993, la cadena de descuento *Target* llegó a un acuerdo fuera del tribunal en una demanda colectiva presentada en beneficio de un estimado de 2 500 solicitantes de empleo. A los futuros guardias de seguridad de *Target* se les pedía que tomaran la *Rodgers Psychscreen* (Psicodetección Rodgers), una combinación condensada de 704 reactivos del CPI y del MMPI. Varios solicitantes estuvieron en desacuerdo para responder la prueba, la cual incluía preguntas sobre Dios, sexo y movimientos intestinales. *Target* acordó pagar 1.3 millones de dólares, incluyendo $60 000 a cuatro demandantes nombrados en la demanda. Aunque *Target* no admitió alevosía en el caso, los funcionarios de la corporación aceptaron no utilizar la prueba Psicodetección durante, cuando menos, cinco años.

Sibi Soraka fue uno de los demandantes. Consideraba que las preguntas eran "descabelladas y raras". Afirmó que el efecto acumulativo de responder a las preguntas lo hacía parecer palpablemente enfermo. Añadió: "No se necesita ser Einstein para darse cuenta de que estas preguntas no tienen en realidad injerencia para nuestro mundo y vida de la actualidad, o con toda seguridad para un trabajo que requiere que se camine por todos lados buscando ladrones de tiendas". La corporación *Target* defendió sus prácticas de prueba, señalando que la Psicodetección se utiliza de manera común en la valoración de oficiales dedicados al cumplimiento de la ley. Los abogados de Soraka estuvieron en desacuerdo y citaron la falta de evidencia de que la prueba ayudara a identificar a personas que fueran buen riesgo contra aquellas que son mal riesgo de empleo. Señalaron que se le negó el empleo a cerca de 800 de los 2 500 solicitantes con base solamente en los resultados de la Psicodetección.

Este caso ilustra que la solidez psicométrica de un instrumento no es el único criterio en la selección de una prueba. Además, los usuarios de pruebas deben demostrar que el instrumento es pertinente para su aplicación. Asimismo, deben considerarse las cuestiones de aceptabilidad para los posibles examinandos.

El lector podría sorprenderse de saber que las pruebas para empleo han hecho surgir controversias legales sólo en los últimos 35 años (Arvey y Faley, 1988). Durante este periodo, varias decisiones judiciales definitivas y directivas gubernamentales que han marcado pauta, han ayudado a definir las tendencias legales actuales. Estos hitos se presentan en el cuadro 11–10, comenzando con el Decreto de 1964 sobre derechos civiles, a través de los reglamentos federales de la *Equal Employment Opportunity Commission*, EEOC (Comisión de Oportunidades Equitativas de Empleo), hasta concluir con los casos judiciales y avances legislativos más recientes. Estos hitos se revisarán en orden cronológico.

PRIMEROS CASOS JUDICIALES Y LEGISLACIÓN

Durante la presidencia de Lyndon B. Johnson, el Congreso de EUA aprobó el Decreto de 1964 sobre derechos civiles. Esta primera legislación tuvo un profundo efecto sobre los procedimientos de prueba para el empleo. Además de disposiciones amplias diseñadas para impedir la discriminación en muchos contextos sociales, el Título VII de este Decreto prohíbe que las prácticas de contratación discriminen con base en la raza, color, religión, sexo u origen nacional. El Decreto estableció varios principios generales importantes que se relacionan con las pruebas de empleo (Cascio, 1987):

- El decreto prohíbe la preferencia discriminatoria por cualquier grupo, minoritario o mayoritario.
- El empleador tiene la carga de la prueba de que todos los requisitos para el empleo, incluyendo las puntuaciones de evaluación, se relacionan con el desempeño en el puesto.
- Las pruebas desarrolladas a nivel profesional que se utilicen en la evaluación del personal deben relacionarse con el trabajo.
- Además de prohibir la discriminación abierta y deliberada, la ley prohíbe las prácticas que son equitativas en cuanto a forma, pero discriminatorias en cuanto a operación.

Cuadro 11–10. Principales hitos legales en las pruebas para empleo

1964	*Myart contra Motorola.* Este caso estableció el precedente para que los tribunales atiendan casos sobre pruebas para empleo.
1964	Decreto de los Derechos Civiles. Este decreto prohíbe la discriminación laboral basada en sexo, raza, color, religión u origen nacional.
1966	Pautas de la EEOC. Primeras pautas publicadas sobre prácticas de prueba para empleo.
1971	*Griggs contra Duke Power.* La Suprema Corte dictaminó que los resultados de una prueba para empleo deben tener un lazo demostrable con el desempeño en el puesto.
1973	*EUA contra Georgia Power.* El dictamen fortaleció la autoridad de las pautas de la EEOC para los estudios de validez de las pruebas de empleo.
1975	*Albemarle contra Moody.* Se fortalecen las pautas de la EEOC; se dictamina que las calificaciones subjetivas de supervisores son una base deficiente para validar pruebas.
1976	*Washington contra Davis.* La corte dictaminó que el desempeño en un programa de entrenamiento constituía una base suficiente contra la cual validar una prueba.
1978	Pautas uniformes para la selección al empleo. Estas pautas definieron el impacto adverso mediante la regla de cuatro quintas partes e incorporaron criterios para la validez de los estudios de selección de empleados.
1988	*Watson contra el Fort Worth Bank and Trust.* El tribunal dictaminó que pueden validarse instrumentos subjetivos para empleo como la entrevista; los empleados pueden reclamar que existe un impacto desigual con base en políticas de promoción basadas en entrevistas.
1990	Decreto sobre los Estadounidenses con Discapacidades. Este decreto limita de manera estricta las razones para no contratar a una persona con discapacidad. Una medida es que no pueden aplicarse pruebas médicas antes de ofrecer un empleo.
1991	Decreto de los Derechos Civiles. Este decreto prohíbe las normas de subgrupo para las pruebas de selección de empleados.

- La intención es irrelevante: el demandante no necesita demostrar que la discriminación fue intencional.
- A pesar de las limitaciones anteriores, las pruebas y otros instrumentos de medición relacionados con el trabajo se consideran tanto legales como útiles.

La legislación de 1964 también creó la Comisión de Oportunidad Equitativa de Empleo (EEOC) para desarrollar pautas que definieran los procedimientos imparciales de selección de empleados. Las pautas iniciales, que se publicaron en 1966, fueron vagas. Revisiones posteriores de éstas, incluyendo las *Uniform Guidelines on Employee Selection* (1978) (Pautas uniformes sobre selección de empleados), fueron bastante específicas y los tribunales las han utilizado para ayudar a resolver disputas legales sobre las prácticas de prueba para el empleo (véase la siguiente sección).

En caso de **Myart contra Motorola** en 1964 marcó la primera participación de los tribunales en las pruebas para empleo. Las cuestiones que surgieron de este caso sobresaliente todavía reverberan. Leon Myart era un afroestadounidense que solicitó empleo en una de las plantas de ensamblado de televisores de la empresa Motorola. A pesar de que tenía experiencia muy pertinente para el empleo, al señor Myart se le negó un puesto porque su calificación en una prueba breve de inteligencia cayó por debajo de la puntuación de corte de la empresa. Myart solicitó una apelación a la Comisión de Prácticas Imparciales de Empleo de Illinois afirmando que existía discriminación racial. El examinador del estado falló en favor del demandante y determinó que la compañía Motorola debería ofrecerle un empleo al señor Myart. Además, el examinador dictaminó que la prueba en particular no debería utilizarse en el futuro y que cualquiera nueva debería "tomar en cuenta los factores ambientales que contribuyen a la privación cultural". En esencia, el examinador concluyó que las prácticas de evaluación de empleo de Motorola eran injustas porque actuaban como una barrera para el empleo de solicitantes con privación y desventajas culturales. Aunque posteriormente se anuló el caso debido a falta de evidencia, **Myart contra Motorola** estableció el precedente para dar audiencia a dichas demandas en el sistema judicial (Arvey y Faley, 1988).

ADVENIMIENTO DE LAS NORMAS DE EVALUACIÓN PARA EL EMPLEO DE LA EEOC

Durante el decenio de 1970, varios casos judiciales ayudaron a moldear las normas y prácticas recientes en cuanto a las pruebas para empleo. El punto central del caso de **Griggs contra Duke Power Company** (1971) era la utilización de pruebas —en este caso la *Wonderlic Personnel Evaluation* y la *Bennett Mechanical Comprehension Test*— como criterios de elegibilidad para empleados que deseaban una transferencia a otros departamentos. En particular, los empleados de la *Duke Power Company* que carecían de una educación secundaria podían estar en condiciones de lograr la transferencia si obtenían puntuaciones por encima de la mediana nacional en ambas pruebas. La política parecía discriminar contra empleados afroestadounidenses, dado que era extremadamente difícil que alcanzaran la elegibilidad para la transferencia. No obstante, los tribunales menores no encontraron intenciones discriminatorias y, por tanto, fallaron en favor de la compañía eléctrica.

En 1971, la Suprema Corte revirtió la decisión del tribunal menor y dictaminó contra la utilización de pruebas sin su validación. La decisión enfatizó varios puntos de relevancia actual (Arvey y Faley, 1988):

- La equidad en las pruebas para empleo se determina por las consecuencias, no por las motivaciones.
- Las prácticas de prueba deben tener un lazo demostrable con el desempeño en el puesto.
- El empleador tiene la carga de la prueba para demostrar que una práctica de empleo como la evaluación con pruebas tiene relación con el puesto.

- Los diplomas, títulos o instrumentos amplios de examen no son medidas adecuadas de la capacidad relacionada con el puesto.
- Las normas de prueba de la EEOC merecen adhesión considerable de parte de quienes hacen evaluaciones para empleo.

Estas pautas para las pruebas de empleo se refinaron todavía más gracias a una decisión judicial de 1973, **EUA contra la Georgia Power Company**. En este caso, la compañía eléctrica de Georgia presentó un estudio de validación para sustentar sus prácticas de prueba para empleo cuando se mostró que sus políticas tenían un impacto adverso en la contratación y transferencia de trabajadores afroestadounidenses. No obstante, el estudio de validez era débil, en parte debido a que se basaba en un análisis de discriminación múltiple, una técnica estadística compleja que se utiliza en pocas ocasiones para este propósito. Los tribunales dictaminaron que el estudio de validación era inadecuado ya que no se adhería a las pautas de la EEOC en cuanto a la evaluación de los estudios de validez. Esta conclusión asentó la posición de las pautas de la EEOC casi como si fuera derecho común en cuanto a las prácticas de prueba para empleo.

Varios otros casos judiciales durante los decenios de 1970 y 1980 también sirvieron para fortalecer la autoridad de las pautas de prueba de la EEOC. Estos casos son bastante complejos e implicaron múltiples cuestiones además de aquellas que se citan aquí. En el caso de **Albemarle contra Moody** (1975), la Suprema Corte se adhirió a las pautas de la EEOC al dictaminar que las calificaciones subjetivas de supervisión son ambiguas y, por tanto, constituyen una base deficiente para evaluar la validez de una prueba de selección para empleo. La cuestión central en el caso de **Washington contra Davis** (1976) era la duda de si el desempeño en un programa de entrenamiento (al contrario del desempeño en el trabajo) era una base suficiente para determinar la relación que pudieran tener con el trabajo los procedimientos de selección para el empleo. En este caso, la Suprema Corte dictaminó que el desempeño en un programa de entrenamiento para oficiales de policía era un criterio suficiente contra el cual validar una prueba de selección.

En el caso del **Estado de Connecticut contra Teal**, la Suprema Corte de EUA se puso en favor de cuatro empleados estatales afroestadounidenses que habían reprobado una prueba escrita que se utilizaba para seleccionar a los solicitantes para el puesto de supervisor de elegibilidad para beneficencia. Los trabajadores afirmaron que existía discriminación injusta, señalando que sólo 54% de los solicitantes de minorías aprobaban el examen, en comparación con 80% de las personas de raza blanca. En su defensa, el estado de Connecticut afirmó que no existía discriminación, dado que 23% de los solicitantes afroestadounidenses exitosos finalmente recibían ascensos, en comparación con 14% de los de raza blanca. El tribunal no se mostró impresionado con este argumento y señaló que el Título VII del Decreto de 1964 sobre derechos civiles se diseñó específicamente para proteger a los individuos y no a los grupos. Así, cualquier falta de equidad hacia un individuo es inaceptable. En Arvey y Faley (1988), Cascio (1987), Kleiman y Faley (1985) y Russell (1984) puede encontrarse un mayor análisis de los casos judiciales sobre empleo equitativo.

PAUTAS UNIFORMES PARA LA SELECCIÓN DE PERSONAL

Durante el decenio de 1970, varias instancias federales y grupos profesionales propusieron revisiones y extensiones para las pautas existentes de la EEOC en cuanto a pruebas para empleo. Las revisiones se desarrollaron en respuesta a las decisiones judiciales que habían interpretado las pautas de la EEOC de manera estrecha, inflexible y legalista. Sin embargo, la existencia de varios conjuntos de pautas competitivas producía confusión y se ejercieron fuertes presiones sobre las partes implicadas para llegar a un acuerdo. Estos esfuerzos culminaron en un documento de consenso conocido como las **Pautas Uniformes para la Selección de Personal** de 1978.

Las pautas uniformes pronto obtuvieron respeto en los casos judiciales y con frecuencia se les cita en resoluciones de disputas legales. Las nuevas pautas contienen interpretación y orientación que no se encuentra en las versiones anteriores, en particular en cuanto al impacto adverso, la equidad y la validación de los procesos de selección, como se analiza más adelante.

Las Pautas uniformes proporcionan una definición muy específica del impacto adverso. En general, cuando los procedimientos de selección favorecen a los solicitantes de un grupo (en general varones o de raza blanca), se dice que la base para la selección tiene un **impacto adverso** sobre otros grupos (en general mujeres o de raza no blanca) con una proporción más baja de selección. Las pautas uniformes definen el impacto adverso con una regla de cuatro quintas partes. En términos específicos, existe un impacto adverso si un grupo tiene una tasa de selección menor a cuatro quintas partes de la del grupo con aquélla más elevada de selección. Por ejemplo, considérese a un empleador que tuvo 200 solicitantes en un año, 100 afroestadounidenses y 100 de raza blanca. Si se contrató a 120 personas, incluyendo 80 de raza blanca y 40 afroestadounidenses, entonces el porcentaje de personas de raza blanca contratada es de 80% (80/100), mientras que el de afroestadounidenses contratados es de 40% (40/100). Dado que la tasa de selección para los afroestadounidenses es de la mitad de aquélla de los de raza blanca (40%/80%), el empleador podría estar vulnerable a una acusación de impacto adverso. Se debe señalar que las pautas uniformes sugieren precaución acerca de esta regla cuando los tamaños de la muestra son pequeños.

Las pautas uniformes también prestan mayor atención a la equidad que los documentos anteriores. Ésta se trata de la siguiente manera:

> Cuando los miembros de un grupo racial, étnico o sexual obtienen de manera característica calificaciones menores en un procedimiento de selección que los de otro grupo y las diferencias no se reflejan en una medida del desempeño en el trabajo, la utilización del procedimiento de selección puede negar de manera injusta oportunidades a los miembros del grupo que obtiene las calificaciones más bajas. Lo que es más, en casos donde dos o más procedimientos de selección tienen igual validez, el empleador está obligado a utilizar el método que produzca el menor impacto adverso.

Las pautas uniformes también establecen una fuerte responsabilidad de discriminación positiva de parte de los empleadores. Si un empleador encuentra una disparidad sustancial en las personas contratadas de un subgrupo en comparación con su disponibilidad en el mercado de trabajo, se recomiendan varios pasos correctivos. Éstos incluyen programas especializados de reclutamiento diseñados para atraer a miembros calificados del grupo en cuestión, programas de entrenamiento en el trabajo de modo que las minorías afectadas no queden atrapadas en empleos sin futuro y una reforma de los procedimientos de selección para reducir o eliminar los efectos de exclusión.

Por último, las pautas proporcionan estándares técnicos específicos para evaluar los estudios de validez de los procedimientos de selección de personal. Casi con toda seguridad los tribunales consultarán estas pautas uniformes si los empleados demandan a una compañía debido a la supuesta injusticia en las prácticas de selección de personal. Así, quien no presta atención especial a estos criterios técnicos es un empleador insensato. Por ejemplo, un criterio se relaciona con la utilización de las puntuaciones de desempeño obtenidas durante programas de entrenamiento:

> Cuando se utiliza como criterio el desempeño en el entrenamiento, el éxito en éste debe medirse de modo apropiado y debe demostrarse la relevancia del entrenamiento ya sea al comparar el contenido del programa de entrenamiento con una conducta o conductas laborales importantes para el trabajo o trabajos, o al demostrar la relación entre las medidas de desempeño en entrenamiento y en el trabajo.

Así, la valoración previa a la contratación de los candidatos al puesto en un programa de entrenamiento puede constituir un método válido

de selección de personal, pero sólo si existe un fuerte vínculo entre las demandas de la tarea dentro del entrenamiento y los requisitos del puesto en sí.

Las pautas uniformes contienen muchos otros criterios que no pueden revisarse aquí. Se insta al lector a leer este fascinante e influyente documento que a menudo se cita en casos judiciales sobre discriminación en el trabajo.

IMPLICACIONES LEGALES DE LAS ESTRATEGIAS SUBJETIVAS PARA EL EMPLEO

En muchas corporaciones, los ascensos se basan en el juicio subjetivo de los gerentes con mayor antigüedad. Una práctica común es que uno o más gerentes entrevisten a empleados calificados y que ofrezcan un ascenso al candidato que parece más prometedor. La selección de este candidato se basa por lo común en la estimación subjetiva de factores tales como juicio, originalidad, ambición, lealtad y tacto. Hasta hace poco tiempo, estas estrategias de empleo parecían estar fuera del espectro de las prácticas imparciales de empleo codificadas en las pautas uniformes y otras fuentes.

No obstante, en un caso civil reciente, **Watson contra el Fort Worth Bank and Trust** (1988), la Suprema Corte facilitó a los empleados la posibilidad de probar los cargos de discriminación racial o sexual contra los empleadores que utilizan entrevistas y otros medios subjetivos de evaluación para la selección o promoción de empleados. Se resumen los antecedentes objetivos de este importante caso antes de analizar las implicaciones legales (Bersoff, 1988).

Clara Watson, una empleada afroestadounidense del *Fort Worth Bank and Trust*, fue rechazada para un ascenso al puesto de supervisora cuatro veces seguidas. En cada ocasión, un solicitante de raza blanca recibió el ascenso. Watson obtuvo evidencia que demostraba que el banco nunca había tenido un funcionario o director afroestadounidense, sólo habían tenido un supervisor de esa raza y le pagaba salarios más bajos a los empleados afroestadounidenses que a los de raza blanca equivalentes. Además, todos los supervisores tenían que recibir aprobación de un vicepresidente de raza blanca, varón, antes de que se les promoviera. El banco no debatió el hecho de que tomara las decisiones de contratación y promoción sólo con base en el juicio subjetivo. Cuando un análisis de los patrones de promoción confirmó disparidades raciales estadísticamente significativas, Watson demandó al banco.

Watson tuvo a la mano dos teorías legales para litigar su demanda bajo el Título VII del Decreto de 1964 sobre derechos civiles. Las dos teorías se denominan "tratamiento desigual" e "impacto desigual". Un caso de tratamiento desigual es más difícil de litigar, dado que el demandante debe probar que el empleador participó en una discriminación intencional. En un caso de impacto desigual, la intención es irrelevante. En vez de ello, el demandante sólo necesita demostrar que una práctica particular de empleo —como la utilización de una prueba estandarizada— resulta en un impacto que es adverso de manera innecesaria y desproporcionada sobre una minoría protegida.

Los tribunales menores dictaminaron que Watson tenía que restringirse al planteamiento más limitado del tratamiento desigual, dado que el empleador había utilizado procedimientos subjetivos de valoración. Además, estos tribunales dictaminaron que el banco no había participado en una discriminación intencional y tenía razones legítimas para no promover a Watson. No obstante, la Suprema Corte estuvo de acuerdo en atender el caso a fin de determinar si podía aplicarse un análisis de impacto desigual a las estrategias subjetivas de empleo, como la entrevista. Basada en gran medida en un informe de la *American Psychological Association* (APA, 1988), la Suprema Corte dictaminó de manera unánime que el análisis de impacto desigual puede aplicarse a las prácticas subjetivas o discrecionales de promoción que se basan en entrevistas. En efecto, el tribunal determinó que pueden validarse las estrategias subjetivas de empleo como la entrevista. Por tanto, los empleadores no tienen la posibilidad de discrecionalidad sin vigilancia para evaluar

a los solicitantes de un ascenso con base en una entrevista subjetiva. Como consecuencia del caso de **Watson contra el Fort Worth Bank and Trust**, los empleadores deben estar listos para defender todas sus prácticas de promoción —incluyendo la entrevista subjetiva— contra demandas por impacto adverso.

AVANCES RECIENTES EN LA SELECCIÓN DE PERSONAL

En 1990, el Congreso de EUA aprobó el Decreto sobre los estadounidenses con discapacidades (ADA), que prohíbe la discriminación contra individuos calificados que tengan una discapacidad. El ADA se analizó de manera breve en el tema 7A, Examen de poblaciones especiales. Este decreto protege a los solicitantes de empleo que tienen discapacidades al limitar de manera estricta las razones permisibles para negarse a contratarlos. En términos específicos, los empleadores pueden negarse a contratar a un trabajador con discapacidad **sólo** por las siguientes razones: 1) si su contratación puede causar dificultades injustificadas a la compañía en cuanto a tener que realizar adaptaciones para la discapacidad; 2) por necesidades comerciales o 3) si la presencia del trabajador con discapacidad puede representar una amenaza directa a la salud o seguridad del trabajador o de los demás.

Una estipulación importante del ADA es que no pueden aplicarse pruebas médicas antes de ofrecer un empleo. Por desgracia, el decreto no define de manera adecuada qué es lo que constituye una "prueba médica". En particular, es posible que se interprete que las pruebas de inteligencia tienen una naturaleza "médica", lo cual podría devastar el campo de las pruebas para empleo:

> Según los requisitos del ADA, si un atributo no se requiere para el desempeño de una tarea esencial, entonces un solicitante puede pedir una adaptación o modificación ya sea del proceso de prueba o del empleo si afirma tener una discapacidad protegida que se asocie con ese atributo no esencial. En la práctica, esto podría significar que, a menos que se demuestre que la inteligencia es un requisito para el logro de una tarea esencial, ninguna prueba que mida inteligencia (o cualquier faceta de ésta) puede aplicarse antes de ofrecer empleo a cualquier solicitante que exprese una alteración asociada con el funcionamiento intelectual (Landy, Shankster y Kohler, 1994).

Aún es demasiado pronto como para determinar el impacto del ADA sobre la práctica de la selección de personal. Existe confusión acerca de cuáles discapacidades cubre el decreto, incertidumbre acerca de cuáles prácticas de evaluación están prohibidas, y ansiedad acerca de cuántas personas buscarán una adaptación con fundamento en el decreto (Klimoski y Palmer, 1994). Se requerirá de decisiones judiciales y pautas administrativas para afinar los detalles de esta importante legislación.

El Decreto de 1991 sobre Derechos Civiles también contiene importantes disposiciones relacionadas con la selección y estimación de personal. En términos específicos, la ley prohíbe normas de subgrupo para las calificaciones de una prueba, lo cual efectivamente elimina la utilización de listas separadas de contratación y promoción. Las normas de subgrupo se refieren a la práctica de utilizar subgrupos identificados (en vez de una muestra nacional diversificada) con el propósito de desarrollar normas de prueba específicas de grupo. La prohibición de esta práctica representa un reto para los empleadores y psicólogos I/O, ya que las normas de subgrupo racial para las calificaciones de una prueba han sido un método popular y eficaz para evitar el impacto adverso.

Los casos judiciales recientes también han tenido repercusiones en las pruebas para el personal. La cuestión en el caso de **Soraka contra Dayton Hudson** fue si las corporaciones pueden utilizar una prueba de personalidad como base para la investigación previa a la contratación en búsqueda de problemas de salud mental en los solicitantes de empleo. Como se analizó antes, se pidió a Soraka que realizara la prueba *Rodgers Psychscreen* (Psicodetección Rodgers) como parte del proceso de solicitud del puesto de guardia de seguridad. La Psicodetección es un inventario de personalidad de falso-verdadero que tiene el propósito de identificar

a personas con problemas psicológicos como depresión y ansiedad. Soraka demandó a la tienda departamental, afirmando que las preguntas individuales acerca de sus prácticas sexuales y creencias religiosas violaban sus derechos civiles. Este caso fue interesante porque tuvo que ver con el valor y validez de los reactivos individuales en contraposición a las calificaciones totales de la prueba. Los tribunales han sostenido durante largo tiempo que las pruebas previas al empleo deben tener una relevancia demostrada para el desempeño en el puesto o no se les puede utilizar. No obstante, los tribunales no han requerido evidencia de validez para los reactivos individuales de prueba. Soraka ganó este caso, que Daytona Hudson apeló. En 1993, la compañía llegó a un arreglo fuera de la corte. Este litigio se resume en el ejemplo de caso 11–2 que se presenta antes en esta sección.

RESUMEN

1. La estimación del desempeño de los empleados es esencial para el éxito continuo de cualquier negocio u organización. La medición psicológica aplicada se coloca al centro de dicha estimación.
2. La valoración del desempeño satisface muchos propósitos de las organizaciones, incluyendo ascensos, transferencias, despidos y establecimiento de salarios. Aunque pueden parecer preferibles los métodos objetivos para la evaluación de la efectividad de los empleados, a menudo los enfoques de juicio son la única opción práctica.
3. Los métodos para la estimación del desempeño incluyen medidas como los cálculos de productividad; datos sobre el personal como la tasa de ausentismo; calificaciones de compañeros y autoevaluaciones; y valoraciones del supervisor como las escalas de calificación. Éstas son, con mucho, el método más común.
4. Cerca de tres cuartas partes de todas las valoraciones del desempeño se basan en métodos de juicio como las escalas de calificación del supervisor. La escala más sencilla es la escala de calificación gráfica, que consiste de etiquetas de rasgos, definiciones breves de dichas etiquetas y de un continuo para la calificación.
5. La escala de clasificación con base conductual (ECBC) es una forma popular de medida del desempeño referida al criterio. Una forma de ECBC contiene referencias conductuales explícitas a lo largo de un continuo de excelencia, que el supervisor evalúa en términos de observaciones pasadas sobre el desempeño en el puesto.
6. La estimación del desempeño está sujeta a diversas fuentes de error que incluyen el fracaso para identificar los criterios apropiados para el desempeño aceptable e inaceptable; el efecto de halo (dar calificaciones altas o bajas a un empleado en todas las dimensiones debido a una impresión global), el sesgo del calificador y la contaminación del criterio.
7. La contaminación del criterio ocurre cuando una medida de criterio incluye factores que no son parte demostrable del trabajo, como calificar a un empleado en cuanto a "trato con el público" cuando esto no es en realidad parte del puesto.
8. Pautas apropiadas para el desarrollo de sistemas de estimación del desempeño incluyen basar el método de estimación en un análisis cuidadoso del puesto; entrenar a los calificadores para que sean imparciales, precisos y legales; y evaluar de manera periódica el sistema de estimación del desempeño.
9. Las pruebas y estimación de los empleados se encuentran cuidadosamente limitadas por

pautas legales y reguladoras. Por ejemplo, el Título VII del Decreto de 1964 sobre derechos civiles prohíbe las prácticas de contratación que discriminen con base en la raza, color, religión, sexo u origen nacional.

10. Varios casos judiciales han ayudado a moldear las prácticas de prueba en la selección de personal. Por ejemplo, en el caso de **Griggs contra la Duke Power** (1971), la Suprema Corte dictaminó que la imparcialidad de las pruebas para empleo se determina por las consecuencias, no por las motivaciones; las prácticas de prueba deben tener un vínculo demostrable con el desempeño en el puesto; y el empleador debe demostrar que las prácticas de prueba se relacionan con el mismo.
11. Varias instancias federales y grupos profesionales ayudaron a desarrollar las *Uniform Guidelines on Employee Selection* (Pautas uniformes para la selección de empleados; 1978). Este documento proporciona orientación acerca de muchas prácticas de evaluación de empleados, incluyendo una definición muy específica del impacto adverso.
12. En general, cuando los procedimientos de selección favorecen a los solicitantes de un grupo (en general varones o de raza blanca), se dice que la base para la selección tiene un impacto adverso sobre otros grupos (en general mujeres o de raza no blanca) cuando tienen una proporción más baja de selección (menos de cuatro quintas partes del grupo de mayoría).
13. Como consecuencia del caso de **Watson contra el Fort Worth Bank and Trust** (1988), en la actualidad los empleadores deben estar en posición de defender todas sus prácticas de promoción —incluyendo la entrevista subjetiva— contra las demandas por impacto adverso.
14. El Decreto de 1991 sobre los estadounidenses con discapacidades también contiene importantes disposiciones relacionadas con la selección y estimación de los empleados. Por ejemplo, el Decreto sobre derechos civiles prohíbe las normas de subgrupo para las pruebas.

TÉRMINOS Y CONCEPTOS CLAVE

Problema del criterio p. 508
Escala de calificación gráfica p. 510
Lista de verificación de incidentes críticos p. 510
Escala de clasificación con base conductual p. 512
Escala de Observación Conductual p. 512
Escala de elección forzosa p. 512
Efecto de Halo p. 514
Sesgo del calificador p. 515
Contaminación del criterio p. 515
Impacto adverso p. 521

■

Capítulo 12

Evaluación vocacional y de valores

Tema 12A
Evaluación de intereses y valores de trabajo

Perspectiva general de la evaluación de intereses
Inventarios para la evaluación de intereses
Evaluación de carrera y de valores de trabajo
Modelo integrador de la evaluación de carrera
Resumen
Términos y conceptos clave

En este capítulo se analizan los enfoques para la evaluación de intereses y valores, definidos de manera amplia. Debido a su valor formativo en todos los aspectos desde el trabajo hasta la devoción, los intereses y valores son fundamentales para la identidad de cada individuo. No es accidental que el adolescente que valora la armonía estética revele posteriormente un interés en la literatura y después siga una vocación como maestro de Español. Tampoco es sorpresa que un adolescente tímido que tiene una tendencia analítica, muestre pasión por las matemáticas y se convierta en un científico informático. Los valores que tienen las personas moldean sus intereses en la vida, los cuales, a su vez, moldean la elección de carrera. La vida posee una congruencia que se explica, en parte, por la influencia de los intereses y valores.

Los valores no sólo vinculan al individuo con el mundo laboral, se encuentran entrelaza-

dos también con las cuestiones morales, espirituales y religiosas. El que se favorezca o se esté en contra de la pena capital, que se considere que la vida tiene significado o es meramente caótica, que se busque o se evite la práctica religiosa —estas cuestiones se deciden con base en los valores personales. En suma, las elecciones que se toman con respecto al trabajo, vida espiritual y conducta personal no son aleatorias, se ligan por los lazos que se denominan intereses y valores.

Un problema que enfrentan muchos jóvenes adultos es que no se han formulado sus valores ni se han explorado sus intereses. Lo que es más, carecen del conocimiento acerca de las opciones de carrera. En estos casos, la selección de carrera puede despertar ansiedad y quizá debería hacerlo. Lowman (1991) ha señalado que el proceso para encontrar una vocación puede ser tan complejo y difícil como elegir una pareja. El dilema de la elección de carrera no se limita a los jóvenes adultos que ingresan al mercado laboral, sino que también inquieta a los trabajadores de mayor edad que se encuentran insatisfechos con sus profesiones. Por fortuna, existe gran cantidad de pruebas y de enfoques de orientación para ayudar a los individuos a identificar sus valores, intereses y elecciones potenciales de carrera, como se analiza en este tema.

En el Tema 12A, Evaluación de intereses y de valores de trabajo, se estudia la medición de aquéllos relacionados con el trabajo. El centro de atención en este caso se coloca en el mundo laboral y en la utilización de pruebas especializadas para la evaluación y orientación de carrera. En el Tema 12B, Valores en el desarrollo moral, espiritual y religioso, se presentan al lector los métodos e inquietudes en la evaluación de las dimensiones moral, espiritual y religiosa del individuo.

PERSPECTIVA GENERAL DE LA EVALUACIÓN DE INTERESES

En la mayor parte de las aplicaciones de pruebas psicológicas, los objetivos de la evaluación son razonablemente claros. Por ejemplo, las pruebas de inteligencia ayudan a predecir el desempeño escolar; las de aptitud predicen el potencial de logro; y las de personalidad proporcionan información acerca del funcionamiento social y emocional. Pero, ¿cuál es el propósito de la evaluación de intereses? ¿Por qué la recomendaría un psicólogo? ¿Qué beneficio puede esperar un cliente de un estudio de este tipo?

La evaluación de intereses promueve dos metas compatibles: la satisfacción vital y la productividad vocacional. Es casi evidente en sí mismo que una buena concordancia entre los intereses individuales y la vocación elegida ayudará a fomentar la satisfacción personal con la vida. Después de todo, cuando el trabajo es interesante es más probable que se experimente también satisfacción personal. Además, las personas satisfechas con su trabajo están en mayor probabilidad de ser productivas. Por tanto, empleados y empleadores por igual obtienen beneficios de la aplicación ingeniosa de la evaluación de intereses. Existen varios instrumentos útiles para este propósito y posteriormente se revisarán los inventarios de intereses de uso más amplio.

En la selección de empleados, la consideración de los intereses personales puede tener gran importancia práctica para los empleadores y, por tanto, también es apropiada de manera contingente para los candidatos al puesto. Puede plantearse una ecuación aproximada como sigue: productividad = capacidad × intereses. En otras palabras, la capacidad elevada en un campo específico no garantiza el éxito, como tampoco lo hace el nivel alto de interés. Es posible tener los mejores medios de predicción cuando se consideran ambas variables. Así, los empleadores tienen una buena razón para determinar si un empleado potencial se adecua bien al puesto; el empleado también debería tener el deseo de saber.

Se comienza esta sección con un examen crítico de las principales pruebas de intereses. Los seis instrumentos elegidos para revisión incluyen los siguientes:

- El *Strong Interest Inventory*, SII (Inventario Strong de Intereses); la última revisión del

muy conocido *Strong Vocational Interest Blank*, SVIB (Formulario Strong de Intereses Vocacionales).
- El *Jackson Vocational Interest Survey*, JVIS (Encuesta Jackson de Intereses Vocacionales), una prueba que incorpora los métodos modernos de construcción de escala.
- El *Kuder General Interest Survey*, KGIS (Estudio de intereses generales de Kuder), un instrumento que incorpora una filosofía divergente de construcción de pruebas.
- El *Vocational Preference Inventory*, VPI (Inventario de preferencia vocacional), que mide seis temas vocacionales ampliamente utilizados.
- La *Self-Directed Search*, SDS (Búsqueda independiente), una guía autoaplicable y autocalificable para explorar opciones de carrera.
- El *Campbell Interest and Skill Survey*, CISS (Encuesta Campbell de intereses y habilidades), una prueba reciente y atractiva que tiene un formato sencillo, pero es compleja en ejecución.

Después de la revisión de las pruebas destacadas de intereses sigue el tema relacionado de la evaluación de valores asociados con la profesión y el trabajo.

INVENTARIOS PARA LA EVALUACIÓN DE INTERESES

INVENTARIO STRONG DE INTERESES

El *Strong Interest Inventory*; SII (Inventario Strong de Intereses) es la última revisión del *Strong Vocational Interest Blank* (SVIB), uno de los instrumentos más antiguos y destacados en las evaluación psicológica (Hansen, 1992). Es posible comprender mejor al SII a través del estudio de la historia de su respetado predecesor, el SVIB. En particular, es necesario revisar las suposiciones que sirvieron de guía para la construcción del SVIB y que se han seguido hasta el SII.

La primera edición del SVIB apareció en 1927, ocho años después de que E. K. Strong formulara los procedimientos esenciales para medir los intereses ocupacionales mientras asistía a un seminario en el *Carnegie Institute of Technology* (Campbell, 1971; Strong, 1927). Para la construcción del SVIB, Strong empleó dos técnicas poco utilizadas en la medición. Primero, se pedía al examinando que expresara su agrado o desagrado hacia una muestra grande y variada de ocupaciones, disciplinas educativas, tipos de personalidad y actividades recreativas. En segundo lugar, se codificaron empíricamente las respuestas para las ocupaciones específicas. En una codificación de este tipo, una respuesta específica (p. ej., agrado hacia el patinaje de ruedas) se asigna a la escala para una ocupación particular, sólo si las personas exitosas en esa ocupación tienden a responder de ese modo con mayor frecuencia que las personas de comparación.

Aunque Strong no expresó sus suposiciones subyacentes de manera simple y directa, es claro que la fundamentación teórica para el SVIB se deriva de una concepción tipológica, orientada a los rasgos, acerca de la personalidad. Tzeng (1987) ha identificado las siguientes suposiciones básicas en el desarrollo y aplicación del SVIB:

1. Cada ocupación tiene un patrón deseable de intereses y características de personalidad entre sus trabajadores. El patrón ideal se representa a través de las personas exitosas en esa ocupación.
2. Cada individuo tiene intereses y rasgos de personalidad relativamente estables. Cuando dichas intereses y rasgos se adecuan a los patrones deseables de intereses de la ocupación, el individuo tiene una probabilidad elevada de entrar a esa ocupación y es más factible que tenga éxito en ella.
3. Es posible diferenciar a los individuos en una ocupación dada de las otras personas en general, en términos de los patrones deseables de intereses y rasgos para esa ocupación.

Strong construyó las escalas de su inventario al contrastar las respuestas de varios grupos cri-

terio de ocupaciones específicas con aquéllas de un grupo de personas en general. Los individuos de cada grupo criterio eran trabajadores en esa ocupación que se sentían satisfechos con sus trabajos y que habían tenido tales empleos durante, cuando menos, tres años. Los reactivos que diferenciaban a los dos grupos, codificados en la dirección apropiada, se seleccionaron para cada escala ocupacional. Por ejemplo, si los miembros de un grupo ocupacional específico sentían desagrado hacia "comprar mercancías para una tienda" con mayor frecuencia que las personas en general, entonces ese reactivo (codificado en la dirección de desagrado) se añadía a la escala para esa ocupación.

El primer SVIB consistió de 420 reactivos y de sólo un puñado de escalas ocupacionales (Strong, 1927). Poco después aparecieron ediciones separadas para varones y mujeres. El inventario ha atravesado por numerosas revisiones a través de los años (Tzeng, 1987), las cuales culminaron en el moderno instrumento conocido como *Strong Interest Inventory* (Inventario Strong de intereses); (Campbell, 1974; Hansen, 1992; Hansen y Campbell, 1985).

Aunque el *Strong Interest Inventory* (SII) se diseñó de acuerdo con la misma filosofía del SVIB, la última revisión se aleja de sus predecesores en tres sentidos cruciales:

1. El SII combina las formas para varones y mujeres en una sola edición.
2. El SII introduce un esquema teórico para guiar la organización e interpretación de las calificaciones, como se analiza más adelante.
3. El SII incorpora un aumento sustancial en el número de escalas ocupacionales, en particular en las áreas vocacionales/técnicas que tenían una baja representación en el SVIB.

El SII consiste de 317 reactivos agrupados en siete secciones. En las primeras cinco, el examinando registra el "Agrado", "Indiferencia" o "Desagrado" por las ocupaciones, materias escolares, actividades, actividades de ocio y contacto con diferentes tipos de personas (cuadro 12–1). Una sexta parte requiere que la persona evaluada exprese una preferencia entre reactivos apareados (p. ej., trato con objetos contra trato con personas). La séptima sección consiste de afirmaciones autodescriptivas que el examinando marca con "Sí", "No", "o"?

Cuadro 12–1. Reactivos característicos del Inventario Strong de Intereses

Marque *Agrado, Indiferencia* o *Desagrado* al lado de los siguientes incisos:

1. Manejar un camión _____
2. Ser funcionario de caza y pesca _____
3. Química _____
4. Realizar investigación aplicada _____
5. Actuar en un drama _____
6. Revistas sobre música _____
7. Sociología _____
8. Obtención de fondos para caridades _____
9. Compra de bienes para una tienda _____
10. Personas que son líderes _____
11. Horario regular de trabajo _____
12. Personas asertivas _____

El SII puede calificarse sólo por medio de hojas de respuesta o protocolos prepagados que se envían por correo o fax a la editorial o a través de la compra de un sistema de *software* que proporciona una calificación en el mismo sitio para resultados inmediatos. Éstos consisten de un impreso extenso que se organiza según diversos temas. Todas las puntuaciones se expresan como puntuaciones estándar con media de 50 y DE de 10. Los resultados normativos para varones y mujeres se informan por separado, pero se pueden lograr comparaciones entre sexos mediante una simple transposición visual.

A nivel más global existen seis calificaciones de tema ocupacional general; a saber, Realista, Investigador, Artístico, Social, Emprendedor y Convencional. Éstas se basaron en el análisis teórico de Holland (1966, 1985ab), cuyo trabajo se analiza después. Cada calificación de tema se asocia con un área de interés principal que describe tanto un ambiente de trabajo como un tipo de persona. Por ejemplo, las personas con

calificaciones altas en el tema Realista son, en general, bastante robustas, tienen dificultad para expresar sus sentimientos y prefieren trabajar en exteriores con maquinaria pesada. Dentro de las calificaciones de tema pueden encontrarse 25 Escalas de intereses básicos como Aventura, Matemáticas y Ciencias Sociales. Las escalas de interés se derivaron empíricamente y consisten de reactivos con intercorrelaciones sustanciales.

Los resultados más específicos consisten de 211 calificaciones para las Escalas Ocupacionales. En la revisión de 1985 del SII, éstas se construyeron de manera habitual al comparar las respuestas de personas empleadas en la ocupación dada contra muestras de varones y de mujeres en general (Hansen, 1992; Hansen y Campbell, 1985). Las magnitudes de muestra para los grupos criterio fueron de 60 a 420 personas, con 200 o más personas en la mayor parte de los grupos. Los grupos criterio consistieron de personas entre los 25 y 60 años de edad, satisfechos con su ocupación, que llenaban ciertas normas mínimas de empleo exitoso y que habían trabajado en la ocupación específica durante, al menos, tres años. La estandarización de la versión de 1985 implicó el examen de más de 140 000 personas, de quienes sólo 50 000 satisficieron los criterios para el desarrollo de escala.

Una innovación reciente del SII es la adición de las escalas de estilo personal (Harmon, Hansen, Borgen y Hammer, 1994). Éstas se diseñaron para medir las preferencias por estilos amplios de vida y trabajo. Estas escalas sirven como auxiliares en la orientación vocacional al mostrar el nivel de comodidad con estilos distintivos. Las cuatro escalas de estilo son:

1. **Estilo de trabajo**, en el que una calificación alta indica preferencia por el trabajo con personas y una baja significa un interés en ideas, datos y cosas;
2. **Ambiente de aprendizaje**, en el que una alta calificación indica preferencia por ambientes académicos de aprendizaje y una baja significa inclinación por actividades más aplicadas de aprendizaje;
3. **Estilo de liderazgo**, en el que una calificación alta indica comodidad para asumir la dirección sobre otras personas y una baja señala intranquilidad; y
4. **Riesgo/Aventura**, en la que una calificación alta indica preferencia por actividades arriesgadas y audaces más que por actividades seguras y fáciles de predecir.

Las escalas de estilo personal tienen una media de 50 y DE de 10. Obsérvese que éstas son escalas realmente bipolares donde cada polo es específico y significativo.

VALORACIÓN DEL SII

El SII representa la culminación de más de 50 años de estudio que implican literalmente miles de informes de investigación y cientos de miles de personas evaluadas. En la valoración de este instrumento, sólo pueden esquematizarse las tendencias básicas en la investigación y referir al lector a otras fuentes para información detallada (Tzeng, 1987; Campbell, 1971; Campbell y Hansen, 1981; Hansen, 1984, 1987, 1992; Hansen y Campbell, 1985). También se debe señalar que las valoraciones de la confiabilidad y validez del SII se basan en parte en su semejanza con el SVIB, para el cual existe una enorme cantidad de datos técnicos.

Con base en estudios *test-retest*, la confiabilidad del SII-SVIB ha resultado ser excepcionalmente buena a corto plazo, con coeficientes de estabilidad a una y dos semanas para las escalas ocupacionales que en términos generales se colocan en el rango del .90. Cuando el intervalo *test-retest* es de años o decenios, las correlaciones descienden hasta los rangos de .60 y .70 para las escalas ocupacionales, excepto en el caso de personas que eran mayores (de 25 años o más) en la primera prueba. Para personas más jóvenes examinadas en la adolescencia, la correlación mediana *test-retest* después de 15 años es de alrededor de .50 (Lubinski, Benbow y Ryan, 1995). Pero para los individuos de mayor edad, examinados por primera vez después de los 25 años de edad, la correla-

ción mediana *test-retest* 10 a 20 años después es un fenomenal .80 (Campbell, 1971). En apariencia, para el momento que se pasa la adultez joven, los intereses personales se vuelven sumamente estables. Las preguntas del SII-SVIB capturan esa estabilidad en las puntuaciones ocupacionales, lo cual proporciona sustentación al concepto de rasgos de personalidad sobre los cuales se basaron estos instrumentos.

La validez del SII-SVIB se funda en gran medida en la capacidad del perfil ocupacional inicial para predecir la ocupación que en un momento dado se practicará. Strong (1955) informó que las probabilidades eran de cerca de dos tercios de que las personas trabajarían en ocupaciones pronosticadas por calificaciones elevadas en las escalas ocupacionales, y de cerca de una de cada cinco de que los individuos examinados estarían en ocupaciones hacia las que hubieran mostrado poco interés cuando se les sometió a prueba. Aunque otros investigadores evitan dar proporciones exactas (Dolliver, Irvin y Bigley, 1972), es claro que el SII-SVIB tiene tasas impresionantes de acierto en la predicción de la ocupación elegida. El instrumento funciona incluso mejor en la predicción de las ocupaciones a las que una persona evaluada **no** entrará. En un estudio reciente, Donnay y Borgen (1996) proporcionan evidencia para la validez de constructo al demostrar fuerte diferenciación general entre 50 grupos ocupacionales en el SII:

> La imagen general es que las personas en diferentes ocupaciones muestran diferencias notables y factibles de predecir en cuanto a agrados y desagrados, ya sea en términos de intereses vocacionales o de estilos personales. Y el inventario Strong proporciona medidas válidas, estructurales y amplias acerca de estas diferencias (p. 290)

El SII se utiliza principalmente con estudiantes de preparatoria y universidad y con adultos que buscan orientación o consejo vocacional para continuar con su educación. Debido a que los intereses de los estudiantes no están desarrollados y son inestables antes de los 13 o 14 años de edad, no se recomienda utilizar el SII antes del nivel de educación media superior. Como es evidente a partir de los datos de confiabilidad que se presentan antes, el SII se vuelve cada vez más valioso con personas mayores y no es poco común observar a aquéllas de mediana edad que utilizan los resultados de este instrumento como orientación para el cambio de carrera.

ENCUESTA JACKSON DE INTERÉS VOCACIONAL

El *Jackson Vocational Interest Survey*, JVIS (Encuesta Jackson de Interés Vocacional), es un instrumento relativamente nuevo que contrasta agudamente en varios sentidos con el SII (Jackson, 1977; Verhoeve, 1993). Las 34 escalas básicas de interés en el JVIS se componen de dos tipos diferentes, escalas de rol de trabajo y escalas de estilo de trabajo. Las 26 escalas de rol de trabajo miden intereses específicos relacionados con temas ocupacionales amplios como matemáticas, ciencias vitales, aventura, comercio y enseñanza. Las ocho escalas de estilo de trabajo se diseñaron para medir preferencias por el trabajo en ambientes que requieren modos particulares de conducta, como seguridad laboral, liderazgo dominante, responsabilidad y vigor. El JVIS puede calificarse manualmente, pero tal vez sea preferible la calificación por computadora, dado que el usuario obtiene entonces varios grupos adicionales de escalas, incluyendo datos sobre la semejanza de la persona evaluada con estudiantes universitarios que se especializan en disciplinas académicas específicas. El JVIS es adecuado para estudiantes de preparatoria y mayores.

Varias características distinguen al JVIS del SII y de otros inventarios de intereses. Primero, el JVIS utiliza un formato ipsativo de opción múltiple donde las personas examinadas deben seleccionar su opción preferida entre dos alternativas. Los reactivos en el JVIS se asemejan a los siguientes:

A. Actuar en una obra teatral escolar.
B. Enseñar a los niños a escribir.

A. Tejer colchas con diseños de ornato.
B. Comprar mobiliario para una cadena de tiendas.

A. Escribir un texto de matemáticas para niños de primaria.
B. Estudiar el desarrollo financiero de un banco local.

Aunque se le utiliza en raras ocasiones, el formato de reactivos de opción forzosa tiene la ventaja de reducir el impacto de la deseabilidad social sobre los resultados de la prueba. Una segunda característica distintiva del JVIS es que Jackson utilizó un método racional y guiado por la teoría para la derivación de las escalas, más que el enfoque empírico que se encuentra en la mayor parte de los demás instrumentos. Como resultado de estos dos aspectos, las escalas del JVIS poseen mayor independencia entre sí en comparación con la que se encuentra en otros instrumentos y también son bastante puras factorialmente. Como evidencia de la homogeneidad de las escalas, las correlaciones biseriales entre reactivos confirmados y puntuaciones de escala se encuentran por lo común en la parte superior del rango de .60 y en la parte inferior del de .70.

Las normas del JVIS se establecieron con una muestra muy amplia —aproximadamente 8 000 estudiantes de preparatoria y universidad. No obstante, estos individuos eran en su mayoría estudiantes de Pennsylvania y de la Provincia de Ontario, de modo que su representación de la población general es cuestionable. La confiabilidad es excelente, al menos a corto plazo, con coeficientes *test-retest* de una a dos semanas que, por lo común, se encuentran a la mitad del rango de .80. Con base en un grupo ecléctico de estudios que se presentan en el **Manual**, la validez concurrente y predictiva parece prometedora, pero se requieren estudios adicionales para reforzar la confianza en este instrumento (Shepard, 1989).

ESTUDIO DE INTERESES GENERALES DE KUDER

El *Kuder General Interest Survey*, KGIS (Estudio de Intereses Generales de Kuder), representa la evolución más reciente en una serie de inventarios de intereses vocacionales de Kuder que han sido sumamente respetados y que se han desarrollado en los últimos 50 años. El primero de estos instrumentos, el *Kuder Preference Record* (Registro de Preferencias de Kuder), se publicó en 1939. Este instrumento introdujo un interesante formato de respuesta de elección forzosa que ha sobrevivido hasta la actualidad (y que se analiza más adelante). El Registro de Preferencias atravesó por diversas revisiones y surgió en 1979 como el *Kuder Occupational Interest Survey –Revised* (KOIS, Kuder y Diamond, 1979) (Estudio de Intereses Ocupacionales de Kuder). El KOIS-R es una prueba muy conocida que produce puntuaciones para más de 100 grupos ocupacionales y casi 50 especialidades universitarias. La población meta para el KOIS-R es aproximadamente la misma que la del SII y del JVIB. Para el propósito de presentar una diversidad de pruebas de interés, es más instructivo analizar el KGIS aquí.

El KGIS es único entre los inventarios de interés en cuanto a que su población meta se restringe a los adolescentes en los grados escolares de bachillerato (Kuder, 1975). La prueba requiere sólo de un nivel de lectura del sexto grado y la puede aplicar el maestro del grupo y calificarla manualmente en el mismo momento. Así, el KGIS es muy adecuado para el desarrollo de metas educativas y vocacionales en los primeros años formativos de la adolescencia.

El KGIS también es poco común en cuanto a su metodología. El inventario utiliza un formato de tríada con respuesta forzosa para medir los intereses. En términos específicos, cada reactivo en la prueba requiere que la persona evaluada indique las alternativas que más le agradan y las que menos de entre tres afirmaciones. Este enfoque de elección forzosa es particularmente adecuado para identificar individuos evaluados que no han respondido con sinceridad a los reactivos.

El inventario con 168 reactivos produce 10 calificaciones de intereses que tienen, en gran medida, una naturaleza ipsativa. El lector recordará que las calificaciones en una prueba ipsativa reflejan variabilidad intraindividual más que interindividual. Con el KGIS, la compara-

ción con un grupo externo de referencia tiene importancia secundaria para determinar las puntuaciones.* Así, una calificación alta en un área de interés significa principalmente que la persona evaluada prefirió con mayor frecuencia dicha área en vez de otras en los reactivos de elección forzosa.

Las 10 escalas reflejan áreas amplias de interés: Exteriores, Mecánica, Computación, Científica, Persuasiva, Artística, Literaria, Musical, Servicio Social y Administrativo. Una undécima escala, la Escala de Verificación, se diseñó para determinar la sinceridad de las respuestas. El manual presenta cuantiosos datos *test-retest*, de consistencia interna y de estabilidad con base en una muestra de 9 819 estudiantes en bachillerato. Los datos de consistencia interna y de *test-retest* a seis semanas son aceptables en términos generales, donde los estudiantes de mayor edad mostraron correlaciones *test-retest* más elevadas. La posible excepción a la buena confiabilidad es la Escala Persuasiva (relacionada con los puestos de ventas), que muestra correlaciones *test-retest* de .69 y .73, respectivamente, para varones y niñas en bachillerato.

Los datos de estabilidad en un seguimiento a cuatro años son menos notables. El coeficiente promedio de estabilidad es de sólo .50 y para personas con CI bajo (menos de 100) es incluso menor, tan bajo como .19 para la Escala Administrativa. Esto es desafortunado ya que los adolescentes con CI bajo están en mayor probabilidad de ingresar a campos administrativos que aquéllos con CI alto. No obstante, la medición de los intereses administrativos es sumamente inestable precisamente para este grupo.

COMENTARIO SOBRE EL KGIS Y OTROS INVENTARIOS DE INTERESES

Si se considera la dificultad de la tarea que se emprende —la medición de los patrones amplios de intereses de los adolescentes— el KGIS tiene un desempeño a nivel aceptable. En bachillerato, los resultados del KGIS pueden alentar a los adolescentes a explorar nuevas experiencias relacionadas con sus intereses medidos; en bachillerato, los resultados pueden ayudar a los estudiantes a planear sus cursos de educación media; y en bachillerato, los resultados pueden ayudar a los estudiantes a tomar decisiones vocacionales tentativas.

Pero el KGIS sufre de las mismas deficiencias centrales de todos los inventarios de intereses existentes, una total falta de atención a la oportunidad. Williams y Williams (1985) han expresado bien esta cuestión:

> ... para aquellos que buscan específicamente una medida de interés, el Kuder es, de manera definitiva, una medida aceptable. Pero el interés es sólo un extremo del triunvirato interés–capacidad–oportunidad. El extremo más importante, la oportunidad, ha generado el menor interés psicométrico. El que esto sea así no es sorpresa. La oportunidad es, con mucho, el constructo más difícil de definir; pero aquellos que trabajan en orientación de carrera nunca deben ignorarlo, sin importar la dificultad en la medición y definición.

Se le recuerda al lector que la falta de atención a la oportunidad es común para todas las medidas de intereses, aunque quizás éste es un problema más grave para el KGIS, debido a que es un instrumento que se utiliza con personas que aún no han ingresado en el mercado laboral.

INVENTARIO DE PREFERENCIA VOCACIONAL

El *Vocational Preference Inventory*, VPI (Inventario de Preferencia Vocacional), es un inventario objetivo, de lápiz y papel, para la medición de la personalidad e intereses, que se utiliza en la evaluación vocacional y de carrera (Holland, 1985c). El VPI mide 11 dimensiones que incluyen los seis temas de personalidad-ambientales denominados: Realista, Investigador, Artístico, Social, Emprendedor y Convencional; y cinco dimensiones adicionales de Autocon-

* En realidad, el KGIS es sólo parcialmente ipsativo. Algunas de las tríadas de los reactivos obtienen puntuaciones para más de una escala.

trol, Masculino/Femenino, Estatus, Infrecuencia y Conformismo. Los reactivos de la prueba consisten de 160 títulos ocupacionales hacia los cuales la persona evaluada expresa un sentimiento al marcar **s** (sí) o **n** (no). El VPI es una prueba breve (15 a 30 minutos) y se dirige a personas de 14 años de edad y mayores con inteligencia normal.

Holland propone que los rasgos de personalidad tienden a agruparse en un pequeño número de patrones relacionados vocacionalmente, denominados tipos. Para cada tipo de personalidad existe también un ambiente laboral correspondiente que es el más adecuado para éste. Según Holland, existen seis tipos: Realista, Investigador, Artístico, Social, Emprendedor y Convencional. En ocasiones esto se conoce como el **modelo RIASEC**, en referencia a las iniciales de los seis tipos. Éstos son idealizaciones en las que pocas personas (o ambientes) encajan por completo. No obstante, Holland considera que la mayoría de los individuos tienden a parecerse más a un tipo que a otros. Además, los individuos también muestran menor grado de semejanza con un segundo y tercer tipos.

Se pueden resumir los tipos de personalidad-ambientales como sigue:

- **Realista:** atlético, carece de habilidades verbales e interpersonales y prefiere vocaciones manuales o en exteriores como las de mecánico, granjero o electricista.
- **Investigador:** pensador orientado a la tarea, con actitudes poco convencionales que encaja bien en puestos científicos o académicos como los de químico, físico o biólogo.
- **Artístico:** individualista, evita situaciones convencionales y prefiere metas estéticas.
- **Social:** utiliza sus capacidades sociales para resolver problemas, le agrada ayudar a otras personas y prefiere profesiones relacionadas con enseñanza o ayuda.
- **Emprendedor:** un líder con buenas habilidades de convencimiento, que encaja bien en puestos comerciales y administrativos.
- **Convencional**: conformista y prefiere roles estructurados como el de cajero de banco u operario de computadoras.

Los seis temas en el sistema RIASEC pueden ordenarse en un hexágono con los temas similares lado a lado y los temas diferentes en lados opuestos, como se representa en la figura 12–1.

Los coeficientes de confiabilidad *test-retest* para las seis escalas principales van de .87 a .97. Las normas del VPI se basan en grandes muestras pertinentes de estudiantes universitarios y adultos con empleo de ediciones anteriores del VPI. Las características de la muestra de estandarización no están bien definidas, lo cual hace que las normas sean un tanto difíciles de interpretar (Rounds, 1985).

La validez del VPI se vincula de manera esencial con la validez del modelo hexagonal de Holland (1985a) sobre los intereses vocacionales. Literalmente cientos de estudios han examinado este modelo desde perspectivas diferentes. Aquí se citarán las tendencias y estudios representativos. Se refiere al lector a Holland (1985c) y Walsh y Holland (1992) para mayores detalles.

Varios estudios sobre el VPI han investigado una cuestión clave en la teoría de Holland —

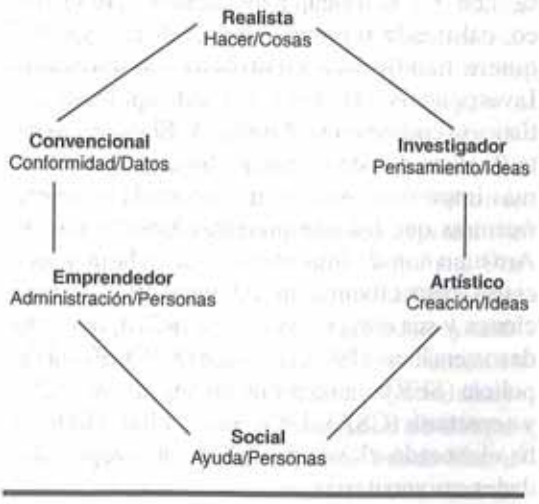


Figura 12–1. Modelo hexagonal de temas ocupacionales de Holland.
Fuente: Reproducido con autorización de Lowman, R. L. (1991). *The clinical practice of career assessment: Interests, abilities, and personality.* Washington, DC: American Psychological Association. Copyright © 1991 de la American Psychological Association.

el que los individuos tienden a dirigirse a ambientes que son congruentes con sus tipos de personalidad. Si esta suposición es correcta, entonces la equiparación en la vida real de ambientes de trabajo y tipos de personalidad de los empleados debería ser sustancial. Debería esperarse encontrar que los ambientes Realistas tienen principalmente empleados Realistas, que los ambientes Sociales tienen principalmente empleados Sociales, y así sucesivamente. La investigación sobre este tema ha seguido una metodología sencilla: se examina a las personas con el VPI y se les clasifica según sus tipos de Holland (utilizando hasta seis letras); después se clasifican de manera independiente los ambientes laborales de los individuos a través de medidas ambientales apropiadas; por último, se calcula el grado de congruencia entre personas y ambientes. En mejores estudios también se aplica una corrección para la concordancia aleatoria.

Al utilizar el modelo hexagonal, Holland ha desarrollado claves ocupacionales como base para clasificar los ambientes de trabajo. Por ejemplo, el arquitecto de paisajes tiene la clave RIA (Realista, Investigador, Artístico), porque se sabe que esta ocupación es un oficio técnico, calificado (componente Realista) que requiere habilidades científicas (componente Investigador) y también demanda aptitudes artísticas (componente Artístico). El componente Realista se lista en primer lugar porque es el más importante para el arquitecto de paisajes, mientras que los componentes Investigador y Artístico son de importancia secundaria y terciaria, respectivamente. Algunas otras ocupaciones y sus claves son taxista (RSE), maestro de matemáticas (ISC), reportero (ASE), oficial de policía (SER), valuador de bienes raíces (ECS) y secretaria (CSA). De modo similar, Holland ha elaborado claves para diferentes especialidades universitarias.

Un enfoque para los estudios de congruencia consiste en comparar los resultados en el VPI de estudiantes o trabajadores con las claves de Holland que corresponden a sus especialidades universitarias u ocupaciones. Por ejemplo, las claves del VPI de Holland para una muestra de oficiales de policía deberían consistir principalmente de perfiles que comienzan con S y debería contener una proporción de perfiles SRE mayor a la debida a factores aleatorios. Lo que es más, el grado de congruencia debería relacionarse con el grado de satisfacción expresada en relación con esa línea de trabajo o de estudio.

La investigación con estudiantes universitarios proporciona fuerte apoyo a la predicción de congruencia: los estudiantes tienden a seleccionar e ingresar a especialidades universitarias que son congruentes con sus tipos principales de personalidad (Holland, 1985a; Walsh y Holland, 1992). Así, los tipos Artísticos tienden a especializarse en arte, los tipos Investigadores tienden a especializarse en biología y los tipos Emprendedores tienden a especializarse en comercio, por citar sólo unos cuantos ejemplos. Estos resultados proporcionan una fuerte sustentación al VPI y a la teoría sobre la cual se basa.

Esta corta reseña apenas ha tocado la superficie de los estudios de validez que apoyan al VPI. Walsh y Holland (1992) citan varias líneas adicionales de investigación que apuntalan la validez de su prueba. Pero no todos los estudios acerca del VPI afirman su validez. Furnham, Toop, Lewis y Fisher (1995) no lograron encontrar relación entre "adecuación" persona-ambiente (P-A) y satisfacción con el trabajo, un fundamento teórico esencial de la prueba. Según la teoría de Holland, entre mejor sea la adecuación P-A, mayor será la satisfacción con el trabajo. En tres muestras británicas, las relaciones eran débiles o inexistentes, lo cual sugiere la posibilidad de que el VPI no se "traslade bien" a culturas fuera de los EUA.

Aunque hasta este momento se han enfatizado principalmente las fortalezas del VPI, incluso los autores de la prueba reconocen que existen posibilidades de mejoría. Por ejemplo, Walsh y Holland (1992) citan las siguientes debilidades del VPI: 1) los conceptos acerca de ambientes vocacionales sólo se han evaluado de manera parcial; 2) las hipótesis acerca de las interacciones persona-ambiente requieren de considerable trabajo adicional de investigación;

3) las formulaciones acerca del desarrollo personal han recibido cierto apoyo, pero necesitan mayores exámenes amplios; 4) la clasificación de las ocupaciones puede diferir dependiendo del instrumento que se utilice para evaluar los tipos de personalidad y 5) existen contingencias personales y ambientales que en la actualidad se encuentran fuera del espectro de la teoría.

La última debilidad es, quizá, la más grave. Después de todo, en la actualidad el enfoque de evaluación del VPI no reconoce ningún papel para la educación, inteligencia y aptitudes especiales, excepto en la medida en que estos factores podrían tener una injerencia indirecta sobre la personalidad y los intereses vocacionales. Sin embargo, el sentido común dicta que la capacidad intelectual tendrá mucho que ver con la satisfacción vocacional en algunas profesiones, independientemente de la adecuación entre tipo de personalidad y ambiente de trabajo. Para un mayor análisis sobre el VPI y la teoría sobre la cual se basa, el lector puede consultar a Gottfredson (1990), Holland (1990) y Holland y Gottfredson (1990).

BÚSQUEDA INDEPENDIENTE

Holland siempre ha mostrado un gran interés en las aplicaciones prácticas de su investigación para el desarrollo vocacional. De acuerdo con este interés, ha desarrollado la Búsqueda Independiente, una prueba breve, sumamente práctica, atractiva por su sencillez (Holland, 1985ab). Como el nombre sugiere, se diseñó para ser una prueba de intereses vocacionales que aplica, califica e interpreta el mismo individuo evaluado y mide los seis temas vocacionales RIASEC que se describieron antes.

Esta prueba consiste en reactivos dicotómicos que la persona examinada marca como "agrado" o "desagrado" (o "sí" o "no") en cuatro secciones: 1) Actividades (6 escalas de 11 reactivos cada una); 2) Competencias (6 escalas de 11 reactivos cada una); 3) Ocupaciones (6 escalas de 14 reactivos cada una), y 4) Estimados personales (2 conjuntos de 6 calificaciones). Para cada sección, los reactivos con validez evidente se agrupan según los temas RIASEC. Para cada tema, el número total de respuestas de "agrado" o "sí" se combina con los estimados personales de capacidad para obtener una calificación total de tema. La prueba requiere de 30 a 50 minutos y se dirige a personas de 15 años y mayores.

Los temas RIASEC en Búsqueda Independiente mostraron confiabilidades *test-retest* que van de .56 a .95 y consistencias internas que van de .70 a .93. Las normas para las escalas y claves de esta prueba se presentan para muestras pertinentes combinadas de 4 675 estudiantes de educación media, 3 355 estudiantes universitarios y 4 250 empleados adultos con edades de 16 a 24 años (Holland, 1985ab). No obstante, los resultados se interpretan, por lo común, de manera individualizada, ipsativa ("¿la ocupación es adecuada para este cliente?"), de modo que los datos normativos tienen relevancia limitada.

Búsqueda Independiente está disponible en una versión de lápiz y papel para calificación manual, y también para computadora. Por desgracia, la versión de lápiz y papel está propensa a una tasa de 16% de errores de calificación cuando la utilizan estudiantes de educación media (Holland, 1985ab). La prueba para computadora personal, práctica para el usuario, es probablemente la versión preferida debido a la facilidad de aplicación y a su calificación e interpretación libre de errores.

Cuando un individuo realiza esta prueba se utilizan las tres calificaciones más elevadas de tema para denotar una clave de resumen. Por ejemplo, una persona cuyas tres calificaciones más altas se colocaron en Investigador, Artístico y Realista, tendría una clave de resumen de IAR. En un folleto independiente que se distribuye con la prueba —el **Buscador de ocupaciones**— la persona examinada puede buscar su clave de resumen y encontrar una lista de ocupaciones más adecuadas. Por ejemplo, un individuo con una clave de resumen IAR encontraría que se asemeja más a las personas en las siguientes ocupaciones: antropólogo, astrónomo, químico, patólogo y físico. El protocolo de prueba contiene información adicional que

ayuda a la persona evaluada a explorar opciones relacionadas de carrera.

Búsqueda Independiente satisface un propósito muy útil al proporcionar un formato rápido y sencillo para instar a las personas jóvenes a examinar sus alternativas profesionales. Al eliminar el proceso lento de aplicación, calificación, interpretación y realimentación del orientador, la prueba posibilita que un público amplio reciba un nivel introductorio de orientación vocacional. Holland (1985ab) propone que esta prueba es apropiada hasta para 50% de estudiantes y adultos que podrían desear orientación vocacional. Supuestamente, para el otro 50%, puede parecer una base insuficiente para la exploración de carrera. Holland (1985ab) advierte de manera correcta a los usuarios que consideren todas las fuentes de información en la elección de profesión, y que no se basen demasiado en las calificaciones de prueba por sí solas. Levinson (1990) analiza la integración de los datos de Búsqueda Independiente con otros datos psicoeducativos para hacer recomendaciones vocacionales específicas a estudiantes de educación media.

La validez de esta prueba se vincula con la validez del modelo hexagonal de la personalidad y los ambientes sobre los que se basa. Por tanto, un aspecto de la validez es si el modelo hace predicciones que se confirmen a través de los resultados de esta prueba en la vida real. En general, los resultados de más de 400 estudios apoyan la validez de constructo de Búsqueda Independiente (Dumenci, 1995; Holland, 1985ab, 1987).

Un planteamiento para determinar la validez consiste en determinar si las relaciones entre las escalas de esta prueba tienen sentido teórico. Como ocurre con el VPI, los seis temas RIASEC de Búsqueda Independiente se pueden ordenar en un hexágono donde los temas similares están juntos y los diferentes se encuentran en lados opuestos. Por ejemplo, en la figura 12–2, los temas Artístico e Investigador se encuentran juntos. No es difícil imaginar a una persona que combine ambos temas en su personalidad y en su ambiente de trabajo, de modo que se puede predecir una correlación positiva moderada entre ellos. En una muestra general de referencia de 175 mujeres entre 26 y 65 años, Holland (1985ab) encontró que las calificaciones en estos dos temas se correlacionaron de manera modesta, $r = .26$, como podría esperarse. El lector también verá que los temas Investigador y Emprendedor se encuentran en lugares opuestos, lo cual significaría enormes disparidades en estos dos patrones ocupacionales. Estos temas no deberían tener correlación. De hecho, las puntuaciones en estos dos temas se correlacionan muy poco, $r = -.02$. En general, las correlaciones que se muestran en la figura 12–2 tienen sentido teórico; estas conclusiones sustentan la validez de constructo de Búsqueda Independiente.

La validez predictiva de esta prueba se ha investigado en varias docenas de estudios que Holland ha resumido (1985ab, 1987). La metodología típica para estos estudios consiste en comparar las claves con las calificaciones más altas de Búsqueda Independiente de grandes muestras de estudiantes con la primera letra de sus elecciones (o aspiraciones) ocupacionales 1 a 3 años después. En general, las conclusiones indican que Búsqueda Independiente tiene una eficiencia de predicción de moderada a alta, dependiendo de la edad de la muestra (las tasas de acierto aumentan con la edad), la extensión del intervalo de tiempo (las tasas de acierto descienden con el tiempo), y la categoría específica pronosticada (las tasas de acierto son mejores para predicciones de los temas Investigador y Social); (Gottfredson y Holland, 1975).

Las correlaciones entre las escalas de Búsqueda Independiente y una amplia variedad de medidas psicológicas (p. ej., personalidad, aptitudes y valores) también sirven para definir el significado de las escalas de esta prueba y, por tanto, ayudan a validar la prueba (Holland, 1985ab, 1987). Por ejemplo, un estudio de Costa, McCrae y Holland (1984) investigó la relación entre las escalas de Búsqueda Independiente y el *NEO Personality Inventory* (Inventario NEO de personalidad) para una muestra de 217 varones y 144 mujeres con edades de 21 a 89 años. Las escalas Investigador y Artístico de Búsqueda Independiente mostraron fuer-

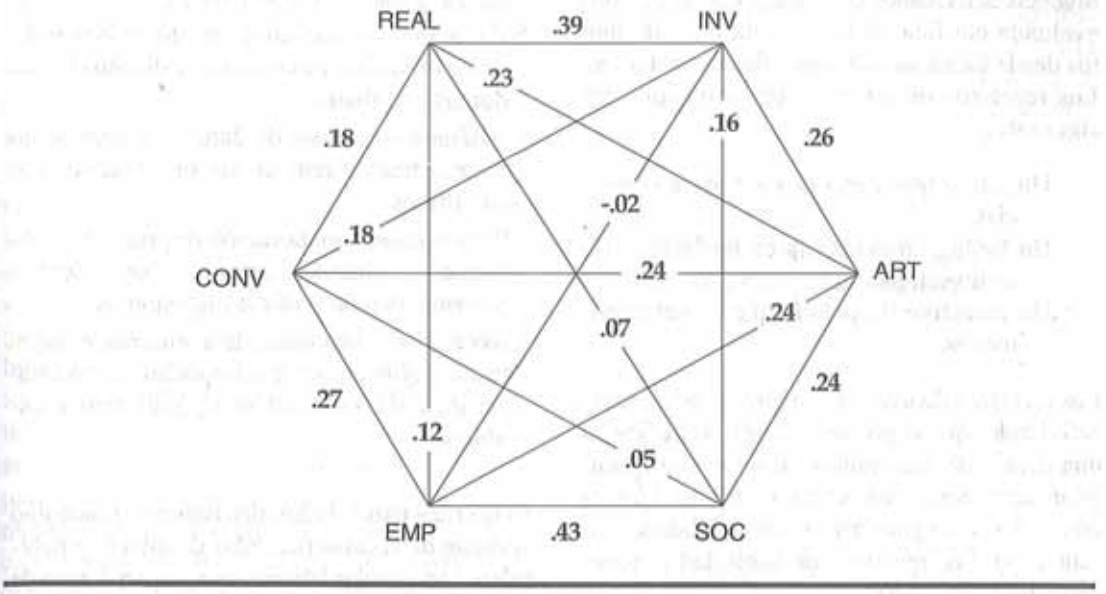


Figura 12–2. Modelo hexagonal SDS y correlaciones entre escalas para 175 mujeres de 26 a 65 años.
Fuente: Adaptado y reproducido con autorización especial del editor del *Self-Directed Search Professional Manual* de John L. Holland, Ph.D. Copyright © 1985 de PAR, Inc. Derechos reservados.

tes correlaciones positivas con la escala de Apertura del NEO —una medida de la apertura a las experiencias en las áreas de fantasía, sentimientos, acciones e ideas. Las escalas Social y Emprendedor de Búsqueda Independiente mostraron fuertes correlaciones con la escala de Extraversión del NEO— una medida de la orientación hacia el exterior y de la sociabilidad. Las escalas Realista y Convencional de Búsqueda Independiente revelaron sólo una correlación trivial con las escalas NEO. En general, las correlaciones observadas fueron consistentes con la interpretación de las escalas I, A, S y E y proporcionan buena sustentación para su validez. Aunque los resultados de este estudio no prestan apoyo a la validez de las escalas R y C, muchas otras investigaciones han producido conclusiones confirmatorias (Holland, 1985ab). Schinka, Dye y Curtiss (1997) proporcionan un concienzudo análisis y discusión sobre la relación entre las dimensiones del NEO y las escalas de la SDS.

ENCUESTA CAMPBELL DE INTERESES Y HABILIDADES

El *Campbell Interest and Skill Survey,* CISS (Estudio Campbell de Intereses y Habilidades; Campbell, Hyne y Nilson, 1992), es una medida más nueva de autoinforme de intereses y habilidades. La prueba se diseñó para ayudar a los individuos a tomar mejores decisiones de carrera al describir la manera en que sus intereses y habilidades se adaptan al mundo laboral. La principal población meta del CISS la conforman estudiantes y jóvenes adultos que no han ingresado al mercado laboral, pero la prueba también es adecuada para trabajadores de mayor edad que están considerando un cambio de carrera. La prueba es apropiada para individuos de 15 años de edad y mayores con un nivel de lectura de sexto grado, aunque en circunstancias excepcionales puede examinarse a niños más pequeños.

La CISS consiste de 200 reactivos de intereses y 120 reactivos de habilidad. Los prime-

ros incluyen ocupaciones, materias escolares y diversas actividades de trabajo que la persona evaluada califica en una escala de seis puntos desde fuerte agrado hasta fuerte desagrado. Los reactivos de intereses se asemejan a los siguientes:

Un piloto que vuela una aeronave comercial.
Un biólogo que trabaja en un laboratorio de investigación.
Un detective de policía que resuelve crímenes.

Los reactivos de habilidad incluyen una lista de actividades que la persona evaluada califica en una escala de seis puntos desde experto (ampliamente reconocido como excelente en esta área), hasta ninguna (no tiene habilidades en esta área). Los reactivos de habilidad se parecen a los siguientes:

Ayudar a una familia a resolver sus conflictos.
Hacer muebles utilizando carpintería y herramientas eléctricas.
Escribir una historia para una revista.

Los resultados del CISS se califican en diferentes tipos de escalas: Escalas de orientación, Escalas de intereses y Habilidades básicas, Escalas ocupacionales, Escalas especiales y Verificaciones de procedimiento. Todas las calificaciones de escala se presentan como puntuaciones *T*, con normas para un promedio de población de 50 y desviación estándar de 10.

Las Escalas de orientación sirven para organizar el perfil del CISS —las escalas de intereses, habilidades y ocupacionales se presenta bajo las Orientaciones apropiadas. Las siete Orientaciones son las siguientes (Campbell y cols., 1992, pp. 2-3):

- **Influencia** —influir en otras personas a través del liderazgo, política, oratoria pública y mercadotecnia.
- **Organización** —organizar el trabajo de otros, administración y vigilancia del desempeño financiero.
- **Ayuda** —ayudar a otros a través de la enseñanza, curación y orientación.
- **Creación** —creación de producciones artísticas, literarias o musicales, y diseño de productos o ambientes.
- **Análisis** —análisis de datos, a través de las matemáticas y realización de experimentos científicos.
- **Producción** —elaboración de productos, utilizando habilidades "manuales" en agricultura, construcción y oficios mecánicos.
- **Aventura** —búsqueda de aventuras, competencia y situaciones que implican riesgo a través de actividades atléticas, policiacas y militares.

Existen 29 pares de Escalas básicas, y cada uno consiste de escalas paralelas de interés y habilidad. Las Escalas básicas se agrupan dentro de siete Orientaciones que se basan en sus intercorrelaciones. Por ejemplo, la Orientación de Ayuda contiene las siguientes Escalas básicas, cada una de las cuales tiene componentes separados de interés y habilidad: Desarrollo adulto, Orientación, Desarrollo infantil, Actividades religiosas y Práctica médica.

Los 58 pares de Escalas ocupacionales, cada una de las cuales tiene componentes separados de interés y habilidad, proporcionan realimentación acerca del grado de semejanza entre la persona evaluada y los trabajadores satisfechos dentro de esa ocupación. Estas escalas se construyeron de manera empírica al contrastar las respuestas de personas que se encontraban satisfechas con su empleo en ocupaciones específicas, con las respuestas de una muestra global de referencia que se obtuvo de la población trabajadora en general.

Además de las Escalas básicas y Ocupacionales, la CISS incorpora tres escalas especiales: Enfoque académico, una medida del interés y la confianza en las actividades intelectuales, científicas y literarias: Extraversión, una medida de la extraversión social; y Variedad, una medida de la amplitud de los intereses y habilidades de la persona evaluada. Por último, el CISS muestra una diversidad de Verifi-

caciones de procedimiento para detectar posibles problemas en la presentación de la prueba, como las respuestas aleatorias o las omisiones excesivas.

En términos generales, la confiabilidad de las escalas del CISS es excepcionalmente fuerte. Por ejemplo, el coeficiente alfa para las Escalas de orientación se coloca por lo común en la parte alta del rango de .80 y las confiabilidades *test-retest* para 324 personas se encuentran en la parte media a alta del rango del .80. Se presentan conclusiones similares sobre confiabilidad para las Escalas básicas y Ocupacionales. Las normas del CISS se basan en 5 000 individuos distribuidos en 58 ocupaciones. Los autores informan de extensos datos de validez para las Escalas ocupacionales, incluyendo medias de muestra para cada muestra ocupacional, así como listas de las tres ocupaciones con calificaciones más altas y más bajas en cada escala (Campbell y cols., 1992). Estos datos documentan que las escalas discriminan entre ocupaciones de una manera eficaz y significativa. Por ejemplo, el promedio de la puntuación *T* en Contabilidad para los contadores es de 75.8. Las personas dedicadas a la estadística, teneduría y planificación financiera logran las siguientes tres calificaciones más altas para esta escala, con puntuaciones *T* promedio en la parte baja del rango de 60. Los artistas comerciales, profesores y trabajadores sociales obtienen las tres calificaciones más bajas, con un promedio en puntuación *T* de alrededor de 40. Estos resultados brindan apoyo a la validez del CISS debido a que se adecuan bien a las expectativas acerca de los patrones ocupacionales de intereses y habilidades.

Casi con toda seguridad este instrumento recibirá mayor atención en los años por venir. Una característica digna de atención en el CISS es la amplitud y claridad del protocolo de informe del perfil. Éste consiste de 11 páginas fáciles de comprender para el usuario. Hemos reimpreso dos de ellas en la figura 12–3 con propósitos ilustrativos. Este formato es preferible a las gráficas detalladas, pero difíciles de leer, que se encuentran en muchos instrumentos. El CISS no tiene aún una fuerte fundamentación de conclusiones empíricas por parte de investigadores independientes. No obstante, la investigación confirmatoria está a la vuelta de la esquina y el CISS promete rivalizar con el *Strong Interest Inventory* para la orientación vocacional de los adultos jóvenes.

EVALUACIÓN DE CARRERA Y DE VALORES DE TRABAJO

En *Working*, su monumental disertación acerca de los estadounidenses en el trabajo, Studs Terkel concluyó que el trabajo es una búsqueda:

> ... de significado cotidiano al igual que del sustento de cada día, de reconocimiento así como de dinero, de asombro en vez de apatía; en pocas palabras, de una especie de vida en lugar de una especie de muerte de lunes a viernes. Quizá la inmortalidad es parte, también, de la búsqueda. Ser recordado fue el deseo, expresado y no expresado, de los héroes y heroínas de este libro (Terkel, 1974).

Las personas buscan significado en su trabajo. Después de entrevistar a cientos de trabajadores, Terkel concluyó que sólo unos cuantos afortunados lo encuentran. Todos podemos recordar a dichos espíritus satisfechos: el ministro religioso que recibe la adoración de su grey, el jardinero paisajista que deja un legado permanente, el mecánico de automotores que se deleita en una maquinaria perfectamente afinada, o el historiador de tradiciones orales que rescata un trozo del pasado. Pero en contraste con estos pocos, Terkel descubrió que la mayoría alberga un "descontento apenas disimulado" acerca de su trabajo.

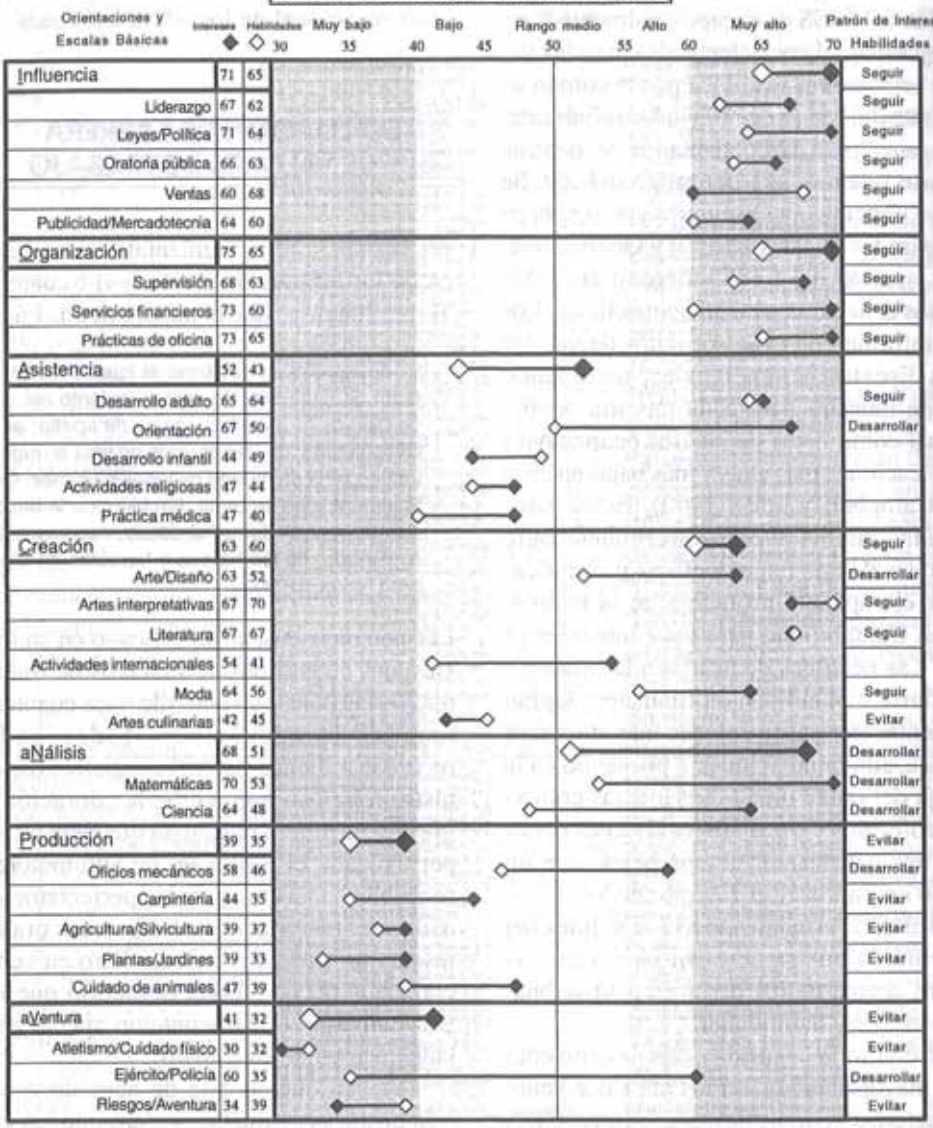
Ya sea que se esté de acuerdo o no con esta postura pesimista, es claro que los valores representan un papel importante en la satisfacción con el trabajo, la elección de carrera y el desarrollo profesional. Esto es especialmente evidente cuando surge una disparidad entre los valores personales y los valores dominantes que requiere una carrera. En su libro sobre cambios de carrera, Jones (1980) relata la historia de un publicista que llegó a una conclusión dolorosa

PERFIL INDIVIDUAL DEL ESTUDIO CAMPBELL DE INTERESES Y HABILIDADES

ESTUDIO DE CASO C

Orientaciones y Escalas Básicas

FECHA DE CALIFICACIÓN: 8/04/92



Nota: El perfil completo consiste de un impreso de 11 páginas.

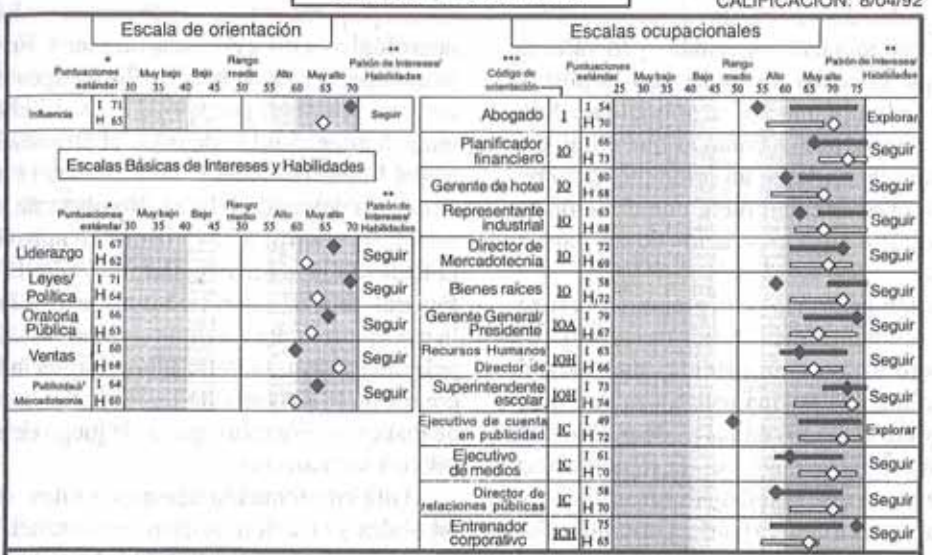
Figura 12–3. Secciones representativas del *Campbell Interest and Skill Survey.*
Fuente: Reimpreso con autorización de Campbell, D. P., Hyne, S. y Nielsen, D. (1992). *Manual for the Campbell Interest and Skill Survey.* Minneapolis, MN: National Computer Systems. Copyright © de David Campbell, Ph. D.

ESTUDIO CAMPBELL DE INTERESES Y HABILIDADES

ESTUDIO DE CASO C

Orientación y Escalas Básicas

FECHA DE CALIFICACIÓN: 8/04/92



La Orientación de Influencia cubre el área general de liderazgo e influencia sobre otras personas. Las personas con altas calificaciones están interesadas en el logro de metas. Desean tomar el cargo y están dispuestas a aceptar la responsabilidad de los resultados. En general, estas personas confían en su capacidad para persuadir a otros de sus puntos de vista y pueden disfrutar de participar en confrontaciones verbales. De manera común trabajan en organizaciones y con frecuencia desean asumir el cargo de actividades específicas que les interesan de manera particular. Disfrutan de hablar en público y les gusta estar a la vista del público. Ocupaciones típicas de personas con altas calificaciones incluyen presidentes de compañías, gerentes de corporaciones y superintendentes escolares.

Usted ha expresado un fuerte interés en el área de liderazgo organizacional —estar a cargo y aceptar la responsabilidad del resultado. Es probable que se sienta cómodo en situaciones donde tiene la responsabilidad de dirigir el trabajo de otras personas, establecer las políticas de la organización y motivar a las personas a su alrededor.

Además, usted también ha presentado un alto nivel de confianza en sus capacidades de liderazgo y motivación a otras personas. Es probable que disfrute estar a cargo de su propio departamento, división u organización y que tenga mucha confianza en que pueda tener un desempeño adecuado. Debido a que sus calificaciones tanto en Intereses como en Habilidades son altas, ésta es un área atractiva para usted.

Sus puntuaciones en las Escalas básicas de influencia, que proporcionan mayores detalles sobre sus Intereses y Habilidades en esta área, se informan arriba en el lado izquierdo de la página. Sus puntuaciones en las Escalas Ocupacionales de Influencia se presentan arriba a mano derecha de la página. Cada ocupación tiene una clave de una a tres letras que indica su calificación o calificaciones más altas en orientación. Entre más similar sea su clave de Orientación a sus calificaciones más altas en Orientación, será más probable que usted se sienta satisfecho con un trabajo en esa ocupación.

* Puntuaciones estándar 1 (✦) = Intereses; H (✧) = Habilidades

** Patrón de Intereses/Habilidades: Seguir = Altos intereses; Alta Habilidad; Desarrollar = Altos intereses, Baja habilidad; Explorar = Alta habilidad; Bajos intereses; Evitar = Bajos intereses, Bajas habilidades.

*** Clave de orientación: I = Influencia; O = Organización; A = Asistencia, C = Creación; N = aNálisis; P = Producción; V = aVentura

Rango de 50% mitad de las personas en esa ocupación; Barra sólida = Intereses; Barra vacía = Habilidades.

Figura 12–3. (Continuación.)

a mitad de la vida: "Me desagrada el punto central de mi trabajo. La publicidad de productos deficientes está dañando al país[...] Todo lo relacionado con la publicidad parecía incorrecto" (p. 27). Esta persona se encontraba tan insatisfecha con su vocación que cambió a otro campo a mitad de su vida. En apariencia, valoraba el servicio hacia los demás —un valor de trabajo que chocaba de frente con la postura amoral predominante en la publicidad. Sólo podemos preguntarnos cómo es que eligió una carrera tan dispar desde un principio, pero parece poco probable que fuera una elección racional basada en una evaluación de sus valores laborales.

También valdría la pena preguntar la fuente del descontento que Terkel descubrió. ¿El descontento es, en gran medida, algo inevitable, inherente a la misma naturaleza del trabajo? ¿O surge, al menos en parte, de millones de discrepancias individuales entre lo que ofrece un empleo y lo que el trabajador necesita?

En esta sección se extenderá la explicación sobre la materia que se desarrolló en el tema anterior —la elección de carrera puede mejorarse a través de la aplicación apropiada de herramientas de evaluación de carrera. El lector encontrará primero las pruebas individuales de valores laborales y desarrollo profesional. A continuación, se analiza el modelo integrador de evaluación de carrera. En este enfoque, se integran las capacidades, intereses y características de personalidad para la orientación vocacional. El capítulo cierra con el tema relacionado de evaluación del consumidor.

Los **valores laborales** se refieren a las necesidades, motivos y valores que influyen la elección vocacional, la satisfacción con el trabajo y el desarrollo profesional. Incluso cuando se mantienen constantes los factores básicos como inteligencia, educación y capacidad, es claro que los individuos difieren en cuanto a sus valores laborales. Aquello que una persona desea de su trabajo podría ser francamente repugnante para otra de igual inteligencia, educación y capacidad.

A continuación se presenta una historia real para ilustrar este punto. En representación de varias familias, un abogado especializado en lesiones personales presentó una demanda en contra de una enorme corporación minera. Se acusaba a la compañía minera de arrojar emisiones venenosas de una fundición de plomo al ambiente que respiraban cientos de residentes de una pequeña población, lo cual causaba daño neurológico sutil a docenas de niños. En la demanda participaron más de 20 testigos expertos y se prolongó por varios años. El abogado tenía fuertes deudas debidas al financiamiento del litigio pendiente —enfrentaba la bancarrota si la demanda fallaba. No obstante, se encontraba jubiloso a medida que se aproximaba la fecha de la confrontación final ante la Corte Federal de EUA, con su futuro profesional en la cuerda floja. Para este individuo, los riesgos peligrosos eran un valor laboral apreciado. En contraste, la mayoría de las personas evitarían de manera activa esta especie de juego de apuestas con sus carreras.

Una combinación apropiada entre valores laborales y elección de carrera es esencial para la satisfacción con el trabajo. Algunas personas tienen éxito en encontrar dicha combinación. Por ejemplo, el intrépido abogado que se menciona antes era muy adecuado para su senda profesional. Pero para aquellos inseguros acerca de la elección de carrera, la realimentación acerca de valores laborales puede proporcionar una orientación muy necesaria. Existen varias herramientas de evaluación que podrían ser útiles en este sentido, pero tres instrumentos merecen atención especial: el *Minnesota Importance Questionnaire*; MIQ (Cuestionario Minnesota de Importancia), el *Work Values Inventory* (Inventario de valores laborales) y la Escala de Valores, que se revisan en la siguiente sección.

CUESTIONARIO MINNESOTA DE IMPORTANCIA

El *Minnesota Importance Questionnaire* (MIQ) se desarrolló para medir las necesidades y valores vocacionales de adultos desde la educación media en adelante. La prueba tiene una

sólida fundamentación en una teoría de adaptación laboral que enfatiza la importancia de la correspondencia persona-ambiente en la determinación de la satisfacción que se deriva del trabajo (Dawis, England y Lofquist, 1964; Dawis y Lofquist, 1984; Lofquist y Dawis, 1991). Según la teoría, la satisfacción con el trabajo se relaciona directamente con la correspondencia entre las necesidades del trabajador y las recompensas o reforzadores disponibles en el trabajo. Por ejemplo, es probable que un futuro empleado que tiene una fuerte necesidad de ayudar a otras personas, encuentre satisfacción en un trabajo que le proporcione amplias oportunidades de servicio social; por el contrario, podría sentirse desdichado en un puesto que enfatice el trabajo solitario.

En su forma más popular —que consiste de reactivos de pares de comparaciones— el MIQ mide 20 necesidades organizadas en seis valores subyacentes relacionados con la satisfacción con el trabajo. La prueba también se presenta en un formato jerarquizado que no se analiza aquí. Los seis valores surgieron de los análisis factoriales de las necesidades. Los valores y sus necesidades componentes se listan en el cuadro 12–2. Es importante enfatizar que cada "escala" de necesidad en realidad consiste de una sola afirmación de necesidad. Por ejemplo, la "escala" de necesidad de Independencia consiste en realidad de una sola afirmación que se asemeja a la siguiente: "Me permitiría tomar mis propias decisiones".

El MIQ consiste de 210 reactivos. Estos incluyen 190 reactivos que aparean cada una de las 30 necesidades con todas las demás. Veinte reactivos adicionales requieren juicios absolutos sobre la importancia de cada dimensión de necesidad. Los reactivos de pares de comparaciones se refieren al "empleo ideal" de la persona evaluada y se asemejan a los siguientes:

_____ Me daría una sensación de logro,
O
_____ Me permitiría tomar mis propias decisiones,

Se instruye a la persona evaluada para que seleccione la alternativa de mayor importancia personal en un empleo —de aquí la referencia al término Importancia en el título del instrumento. En el ejemplo anterior, la necesidad de logro (Me daría una sensación de logro) se equipara con la necesidad de responsabilidad (Me permitiría tomar mis propias decisiones). A fin de aparear cada necesidad con cada una de las otras, se requieren 190 reactivos.* Cada una de las 20 necesidades laborales también se clasifican de manera individual en una escala absoluta de importancia, lo cual resulta en un total de 210 reactivos. Estos juicios absolutos permiten las comparaciones entre personas evaluadas o entre escalas para un mismo individuo examinado.

El MIQ se interpreta en referencia a **patrones de reforzamiento ocupacional** (PRO) para

Cuadro 12–2. Valores y componentes del Cuestionario Minnesota de importancia

Valores	Componentes
Aprovechamiento	Utilización de capacidad Aprovechamiento
Comodidad	Actividad Independencia Variedad Compensación Seguridad Condiciones de trabajo
Estatus	Avance Reconocimiento Autoridad
Altruismo	Compañeros de trabajo Servicio social Sentimientos morales
Seguridad	Políticas de la compañía Supervisión—Relaciones humanas Supervisión—Técnica
Autonomía	Creatividad Responsabilidad

* El número total de comparaciones de pares para N afirmaciones (cuando el orden no es importante) puede calcularse a partir de la fórmula $N(N\text{-}1)/2$ que es $20(19)/2$ o 190.

casi 200 ocupaciones. Los PRO se derivaron de un programa de investigación paralela que utilizó el *Minnesota Job Description Questionnaire*, MJDQ (Cuestionario Minnesota de descripción de puestos), una escala que se parece al MIQ. El MJDQ requiere que los actuales titulares del puesto califiquen la presencia o ausencia percibida de reforzadores en un empleo dado. Por supuesto, éstos son simplemente las 20 necesidades laborales reformuladas en términos apropiados, de modo que capten los requisitos ocupacionales.

Mediante la comparación del perfil de necesidades y valores de la persona evaluada con los patrones conocidos de reforzamiento para las ocupaciones representativas, los resultados del MIQ pueden utilizarse para pronosticar la satisfacción en empleos específicos. Esto se realiza por medio del Índice C o índice de correspondencia, que es el coeficiente de correlación entre el perfil del individuo en el MIQ y el PRO para cada ocupación. Se predice satisfacción con una ocupación cuando la correlación supera al .50.

El formato de pares de comparaciones del MIQ permite que el examinador valore la consistencia de las respuestas. Considérense tres necesidades cualesquiera, designadas como **A**, **B** y **C**. Supóngase que una persona evaluada prefiere **A** por encima de **B** y también prefiere a **B** por encima de **C**. En términos lógicos, esta persona también preferiría **A** por encima de **C** (transitividad). Éste es un ejemplo de una Tríada Lógicamente Consistente (TLC). La puntuación TLC es el porcentaje de todas las tríadas que son lógicamente consistentes. Esta calificación proporciona un índice de la consistencia de las respuestas, que es una medida de la validez de presentación de la prueba. Las puntuaciones TLC por debajo de 33 hacen surgir la sospecha de que la persona evaluada ha respondido de manera descuidada o aleatoria. En los estudios *test-retest*, entre mayor sea la puntuación TLC, más estable será el perfil MIQ de la persona examinada.

La confiabilidad del MIQ va de adecuada a excelente, dependiendo del intervalo para la repetición de la prueba. Se informa que la mediana de la correlación *test-retest* para las 20 escalas es de .89 para una repetición inmediata de la prueba, pero de sólo .53 cuando la repetición de ésta es después de 10 meses. Las confiabilidades de consistencia interna se colocan por lo común alrededor del .80 (Rounds y cols., 1981).

Cerca de 200 estudios tratan sobre la validez del MIQ, de modo que es difícil resumir las tendencias (Layton, 1992). Los resultados indican que las 20 escalas del MIQ discriminan entre distintos grupos ocupacionales; las correlaciones con el *Strong Vocational Interest Blank* son significativas y consistentes con la teoría; y el MIQ tiene correlaciones apropiadamente bajas con las capacidades como las mide la *General Aptitude Test Battery* (Batería de pruebas de aptitud general; Benson, 1985; Layton, 1992). En un estudio confirmatorio, las puntuaciones de Independencia del MIQ predijeron de manera moderada si los estudiantes de posgrado en orientación psicológica se volverían científicos o practicantes cuando ingresaran al mercado laboral (Tinsley, Tinsley, Boone y Shim-Li, 1993).

El MIQ es un instrumento muy respetado que merece utilizarse de manera extensa. Curiosamente, la prueba nunca ha captado en realidad una atención amplia. Quizás esto se debe al formato del instrumento, que podría constituir un impedimento para que lo adopte el personal de servicios humanos. El problema es que las personas evaluadas se enfrentan con las mismas afirmaciones de necesidad una y otra vez. A fin de aparear cada una con todas las demás necesidades, es necesario utilizar cada afirmación individual en 19 reactivos separados de la prueba. Los individuos examinados sienten que se encuentran con las mismas preguntas una y otra vez, aunque cada reactivo del MIQ es, de hecho, único. Sin considerar su solidez psicométrica, desde el punto de vista de la persona evaluada, el MIQ es un instrumento poco agradable.

INVENTARIO DE VALORES LABORALES

El *Work Values Inventory*, WVI (Inventario de Valores Laborales), es un instrumento corto y

sencillo diseñado para medir 15 valores laborales en individuos desde la secundaria hasta la preparatoria (Super, 1968, 1970). La prueba es el producto final de decenios de investigación acerca de las metas que motivan a los individuos a trabajar. Los 15 valores laborales se identificaron a través de una revisión de la literatura que incluyó el primer trabajo clásico de Spranger (1928), sobre los **Tipos de hombres**. Los reactivos, escalas y formatos de la prueba se revisaron y se refinaron continuamente hasta seleccionar el planteamiento actual de calificación de cinco puntos (Super, 1970, 1973).

El WVI es un instrumento de autoinforme que consiste de 45 reactivos calificados en una escala de cinco puntos desde "Muy importante" hasta "Sin importancia". Los reactivos de la prueba se parecen a los siguientes: "Volverse famoso en su campo". "Tomar sus propias decisiones en el trabajo", "Sentir que ha ayudado a otras personas" y "Tener un jefe considerado". Hay tres reactivos en cada una de las 15 escalas. Los 15 valores laborales que mide la prueba incluyen los siguientes:

Altruismo	Remuneración económica
Estética	Seguridad
Creatividad	Contorno
Estimulación intelectual	Relaciones de supervisión
Logro	Colegas
Independencia	Modo de vida
Prestigio	Variedad
Dirección	

Estos valores laborales se describen en detalle en el manual. Por ejemplo, Altruismo se describe como "presente en un trabajo que le permite a uno contribuir al bienestar de los demás" y Prestigio se explica como "asociado con un trabajo que le proporciona a uno una posición a los ojos de los demás y evoca respeto". Por ejemplo, "Volverse valioso en su campo" pertenecería a la escala de Prestigio.

Los resultados del WVI se presentan como 15 puntuaciones naturales de escala, cada una de las cuales va desde 3 (calificación baja) hasta 15 (responder con "Muy importante" a los tres reactivos en la escala). Estas calificaciones permiten tres estrategias de interpretación. En un análisis clínico se destacan las tres calificaciones más altas con propósitos de discusión por parte del orientador y de la persona evaluada. Los datos también pueden analizarse en términos normativos con respecto a los resultados de otros estudiantes de la misma edad. Más importante es el hecho de que es posible predecir la satisfacción en diversas ocupaciones mediante el empleo de los patrones de reforzamiento ocupacional (PRO). El enfoque es similar al enunciado anteriormente para el MIQ, en el que el orientador determina el grado de equiparación entre los valores laborales de la persona examinada y los reforzadores conocidos, disponibles en diversas ocupaciones.

Los aspectos técnicos del WVI son encomiables. En una muestra de 99 estudiantes de bachillerato, el instrumento reveló confiabilidades *test-retest* a dos semanas en el rango de .80 para todas las escalas excepto Prestigio (r = .76). El manual proporciona cuantiosos datos normativos para estudiantes de secundaria y preparatoria. La validez también parece robusta, como se juzga por las correlaciones con otras medidas de valores laborales, análisis factorial de las escalas y reactivos, y relaciones con criterios externos que confirman la teoría. Bolton (1985) proporciona una excelente reseña de la evidencia de validez para el WVI. Quizás el único comentario precautorio acerca de éste es que el instrumento es, en la actualidad, un tanto obsoleto. Es necesario que se proporcionen nuevas normas y datos sobre confiabilidad.

ESCALA DE VALORES

Un consorcio de investigadores bajo la dirección de Super y Nevill (1986) desarrolló la Escala de valores (VS) para evaluar 21 valores relacionados con los roles laborales y vitales. La prueba consiste de cinco reactivos por valor, calificado cada uno desde 1 ("De poca o ninguna importancia") hasta 4 ("Muy impor-

tante"). Se utiliza un reactivo final cuando la escala se aplica de manera transcultural, para un total de 106 reactivos. Los valores que mide la VS incluyen los siguientes:

Utilización de capacidad	Desarrollo personal
Logro	Actividad física
Avance	Prestigio
Estética	Riesgo
Altruismo	Interacción social
Autonomía	Variedad
Creatividad	Condiciones laborales
Recompensas económicas	Identidad cultural
Estilo de vida	Destreza física
	Seguridad económica

Un aspecto poco común, pero sumamente deseable de la VS es que la prueba se desarrolló de manera explícita para utilizarse en investigación transcultural. Un consorcio informal de equipos de investigación de docenas de países en América del Norte, Europa, Australia, Asia y África participó en la definición, revisión y refinamiento de los valores que mide la prueba. Cada equipo nacional la tradujo a su propio idioma y piloteó los reactivos.

La confiabilidad de la VS es meramente adecuada, lo cual es comprensible dado que los instrumentos sólo contienen cinco reactivos por escala. Los coeficientes alfa se encuentran por arriba de .70 en todas las escalas y las confiabilidades *test-retest* se colocan arriba de .50 en muestras de estudiantes universitarios. Se proporcionan normas para una muestra apropiada de 3 000 estudiantes y adultos estadounidenses. Los estudios iniciales de validez parecen prometedores, pero se requieren más antes de que la prueba se emplee para orientación individual (Rousseau, 1989; Slaney, 1989).

La Escala de valores representa los ideales más elevados en el desarrollo de pruebas. Al incluir la participación de equipos de investigación en todo el mundo, Super y Nevill (1986) han concebido una prueba con verdadero atractivo y utilidad transculturales. Quizá sus esfuerzos ayudarán a forjar una perspectiva global sobre la naturaleza y valor del trabajo. Con demasiada frecuencia, el desarrollo de pruebas ha sido una actividad territorial restringida a las culturas industrializadas de occidente. Sólo puede esperarse que otros autores de pruebas también valoren la perspectiva transcultural en la evaluación.

EVALUACIÓN DEL DESARROLLO PROFESIONAL

Super (1957, 1990) ha enfatizado que la elección de carrera no es una decisión separada sino un proceso continuo. Este autor afirma que el desarrollo profesional se caracteriza por etapas: crecimiento, exploración, establecimiento, mantenimiento y declinación. En la etapa de crecimiento, un individuo se entretiene en fantasías, desarrolla intereses y descubre sus capacidades. Durante la etapa de exploración en la adolescencia y adultez temprana, el individuo participa en un examen tentativo de las carreras. A esto le sigue la estabilización y consolidación de una carrera en la fase de establecimiento. Aproximadamente a la edad de 50 años, la mayoría de los individuos entra a la etapa de mantenimiento, la cual se caracteriza por la innovación y actualización para algunos, pero por el estancamiento y la desaceleración para otros. La etapa de declinación se caracteriza por la disgregación para la mayoría de las personas, pero para unas cuantas implica una especialización profesional.

En las etapas iniciales del desarrollo profesional —las etapas de crecimiento y exploración— es posible que las pruebas vocacionales tradicionales no proporcionen los mejores tipos de información para orientación, dado que en general se basan en la premisa de que la persona evaluada está bien informada acerca del trabajo y tiene patrones de interés bien establecidos. No obstante, es común que los individuos en estas etapas tengan información limitada sobre las carreras y conocimientos mínimos de sus intereses y valores vocacionales. En tales

situaciones, se requiere de instrumentos especializados para una evaluación vocacional eficaz.

Varias medidas vocacionales se basan en el reconocimiento de que la elección de carrera es un proceso continuo en vez de una sola decisión. Estos instrumentos alternativos se enfocan en la madurez del conocimiento sobre la carrera, la planificación vocacional y las habilidades en la toma de decisiones. En el cuadro 12–3 se mencionan brevemente varias pruebas representativas acerca del desarrollo de carrera y la madurez vocacional. Para un análisis más amplio, se insta al lector a consultar a Walsh y Betz (1995).

Cuadro 12–3. Medidas representativas del desarrollo profesional

Career Directions Inventory **(Inventario de Orientación Profesional)**

(Jackson, 1986)

Consiste de 100 tríadas de afirmaciones que describen actividades relacionadas con el trabajo y el examinando marca sus actividades más preferidas y menos preferidas. Las 15 escalas de intereses básicos incluyen tanto roles laborales como estilos de trabajo; p. ej., administración, servicios alimenticios, ventas, exteriores, literatura, asertividad, persuasión, sistematización. Excelentes confiabilidad y validez; las normas se basan en 12 000 individuos de más de 150 especialidades educativas y ocupacionales.

Escala de Decisiones Profesionales

(Osipow, 1980)

Inventario corto que consiste de 19 reactivos en los que la persona evaluada indica en una escala de 1 a 4, qué tan estrechamente describe cada una de las afirmaciones su pensamiento acerca de las profesiones. Esta escala proporciona una medida de la indecisión sobre la carrera que es de particular utilidad en los estudios previos y posteriores a las intervenciones vocacionales.

Career Beliefs Inventory **(Inventario de Creencias sobre la Profesión)**

(Krumboltz, 1991)

Formado por 96 reactivos que se califican en una escala de cinco puntos desde "fuertemente de acuerdo" hasta "fuertemente en desacuerdo", este inventario tiene el propósito de identificar las creencias del cliente que pueden obstaculizar el logro de metas profesionales. Ejemplos de las 25 escalas incluyen: planes de carrera, aceptación de la incertidumbre, satisfacción intrínseca, control, aprobación de los demás.

Career Development Inventory **(Inventario de Desarrollo Profesional)**

(Super, Thompson, Lindeman y otros, 1981)

El CDI es una medida amplia del desarrollo y madurez profesional que consiste de cinco subpruebas: Planificación de carrera, que mide el grado y participación en ésta; Exploración de carrera, que valora los intentos actuales y futuros para obtener información sobre la carrera; Toma de decisiones, que mide la capacidad para aplicar el conocimiento y discernimiento a la planificación de carrera; Información del mundo laboral, una medida del conocimiento acerca de la estructura ocupacional; y Conocimiento sobre el grupo ocupacional preferido, que proporciona una evaluación a profundidad del conocimiento acerca del grupo ocupacional preferido de la persona evaluada.

MODELO INTEGRADOR DE LA EVALUACIÓN DE CARRERA

Es poco común que los profesionales de la evaluación de carrera dependan de la información de una sola fuente como un estudio de intereses o valores laborales. El orientador vocacional eficaz utiliza un **modelo integrador** en el que se considera de manera simultánea la información de las áreas de intereses, capacidad y personalidad. Lowman (1991) ha presentado los elementos de este enfoque y gran parte del siguiente análisis se basa en su análisis. Los profesionales de este método no minimizan la importancia de los intereses en la determinación de la elección de carrera y la satisfacción con la misma. No obstante, tienden a asignar

una importancia primaria a los patrones de capacidad y a ciertas características de personalidad en la evaluación y orientación vocacional. En primer término se analizan los patrones de capacidad.

PATRONES DE CAPACIDAD EN LA EVALUACIÓN DE CARRERA

Dependiendo de las metas vocacionales del cliente, varias dimensiones de capacidad podrían ser pertinentes para la evaluación de carrera. Una lista parcial incluye capacidades físicas, análisis espacial, inteligencia verbal, capacidades de investigación, capacidades artísticas e inteligencia social (Gottfredson, 1986; Lowman, 1991). La importancia de las capacidades mentales amplias o primarias como el análisis espacial o la inteligencia verbal es bastante obvia. Por ejemplo, la arquitectura requiere de niveles elevados de análisis espacial para lograr el éxito. En el caso de un futuro arquitecto, ninguna cantidad de interés en el campo puede compensar una baja capacidad en análisis espacial. De la misma manera, las capacidades verbales son esenciales para el periodismo y otras profesiones que demandan dominio del lenguaje. La importancia de aptitudes especializadas como las capacidades mecánicas (para un futuro mecánico) y las capacidades artísticas (para un futuro artista) es, de la misma manera, algo explícito. Pero, ¿qué hay con respecto a la inteligencia social? ¿Es pertinente para la evaluación de carrera? ¿Se puede medir la inteligencia social?

Thorndike (1920b) fue el primero en identificar la **inteligencia social**, la cual se refiere a la capacidad para comprender a otras personas y relacionarse de manera eficaz con ellas. Aunque ha habido una controversia continua acerca de la validez del constructo de la inteligencia social, estudios recientes indican que pueden utilizarse medidas sencillas de lápiz y papel para aislar esta dimensión de capacidad de otros aspectos de la inteligencia (Lowman y Leeman, 1988). Se revisarán dos estudios para ilustrar esta cuestión.

Getter y Nowinski (1981) desarrollaron la *Interpersonal Problem Solving Assessment Technique*, IPSAT (Técnica de Evaluación para Solución de Problemas Interpersonales), una prueba semiestructurada de respuesta libre acerca de la efectividad interpersonal. En esta prueba se presenta a la persona evaluada una serie de 46 situaciones interpersonales problemáticas y se le pide que se imagine dentro de cada una de ellas. Se instruye a los examinandos a escribir maneras alternativas de manejar cada situación y se les pide que indiquen cuál de estas soluciones potenciales elegirían de hecho. Un ejemplo de una situación es como sigue:

> Tu jefe (o maestro) acaba de criticar una parte del trabajo que has realizado y consideras que la crítica es injustificada e injusta. ¿Qué harías?

Con base en un manual detallado de calificaciones, se califica cada una de las respuestas elegidas por la persona evaluada en una de estas categorías: Efectiva, Evitante, Inapropiada, Dependiente e Incalificable. Primero se calcula el número total de respuestas para proporcionar un índice de la capacidad de la persona evaluada para pensar en cursos alternativos de acción. Después, se calcula el número de soluciones elegidas calificadas dentro de cada categoría, lo cual proporciona un perfil de los tipos de soluciones que prefiere el individuo examinado. La confiabilidad intercalificadores de las subescalas de la IPSAT es bastante alta y las correlaciones con otros instrumentos prestan un fuerte apoyo para la validez convergente y de discriminación de este instrumento.

Un inventario más reciente y prometedor de la inteligencia social es el *Social Relations Survey*, SRS (Encuesta de Relaciones Sociales), un instrumento con 128 reactivos de falso-verdadero que desarrollaron Lorr, Youniss y Stefic (1991). Estos autores utilizan un método de construcción de escala racional que se sostiene en el análisis factorial para producir un instrumento que mide ocho factores de la inteligencia social. En el cuadro 12–4 se presentan las subescalas y reactivos ilustrativos. Para los 49 participantes que volvieron a examinarse dos semanas después, la confiabilidad

Cuadro 12–4. Escalas y reactivos ilustrativos de la Encuesta de relaciones sociales

Asertividad social

Me resulta fácil hablar con otras personas a las que acabo de conocer. (V)

Cuando conozco a nuevas personas, generalmente dejo que elijan el tema de la plática. (F)

Directividad

Estoy en mi mejor momento cuando soy la persona que lleva el control. (V)

Me siento cómodo cuando consiento en que otros tomen la dirección de un grupo. (F)

Defensa de derechos

Si alguien rebasa mis límites, me defiendo. (V)

Me incomoda regresar mercancía a una tienda. (F)

Confianza

Me siento confiado la mayor parte del tiempo. (V)

Me siento insatisfecho de mis capacidades. (F)

Aprobación percibida

Estoy seguro de que las personas que conozco piensan bien de mí. (V)

En ocasiones tengo una sensación de desaprobación de las personas a mi alrededor. (F)

Expresión de sentimientos positivos

Me agrada mostrar mis sentimientos positivos hacia los demás. (V)

Me incomoda mostrar afecto por un amigo en público. (F)

Necesidad de aprobación social

Hago esfuerzos deliberados por volverme popular. (V)

No me preocupa lo que otras personas digan de mí. (F)

Empatía

Me siento sumamente afectado cuando mis amigos me cuentan sus problemas. (V)

En general conservo sentimientos objetivos e imparciales hacia los demás. (F)

Fuente: Basado en Lorr, M., Youniss, R., y Stefic, E. (1991). An inventory of social skills. *Journal of Personality Assessment, 57*, 506-520.

test-retest mediana de las subescalas fue un impresionante .89. Se proporcionan normas para 260 mujeres universitarias y 355 mujeres con educación media. Varios enfoques para la validez concurrente y de constructo indican que el SRS proporciona un abordaje válido para la evaluación de autoinforme de las habilidades sociales.

Más allá de toda duda, la inteligencia social está sumamente relacionada con la orientación de carrera. Por ejemplo, una futura enfermera requerirá de altos niveles de inteligencia social para funcionar de manera eficaz en ese puesto. En contraste, un técnico de computadoras puede necesitar muy pocas habilidades sociales para tener un desempeño superior en el ambiente de trabajo. Lowman (1991) presenta una taxonomía hipotética de la inteligencia social para ilustrar su relevancia en la evaluación de carrera (cuadro 12–5). Aunque la medición de la inteligencia social continúa siendo un reto, los profesionales del área harían mal en ignorar este factor de capacidad en la evaluación de carrera.

PATRONES DE PERSONALIDAD EN LA EVALUACIÓN DE CARRERA

Varios patrones de personalidad también son sumamente pertinentes para la evaluación de carrera. Las pruebas de personalidad se analizan en detalle en capítulos posteriores, por lo cual sólo se mencionarán aquí unas cuantas dimensiones de personalidad laboralmente pertinentes.

- La necesidad de logro es importante para las personas con aspiraciones en profesiones de negocios y de dirección (p. ej., Orpen, 1983).
- La ascendencia o dominio también es importante para la jerarquía de dirección (p. ej., Bentz, 1985).
- La estabilidad emocional predice el desempeño positivo en una amplia gama de trabajos tradicionales, mientras que el neuroticismo se asocia con el éxito en algunas profesiones artísticas (p. ej., Wills, 1984).

Cuadro 12–5. Taxonomía hipotética de las demandas sociales de los puestos

Grado de participación social	Dimensiones sociales del puesto
Muy alto	Roles terapéuticos, educativos o administrativos, como gerente comercial, enfermera o psicoterapeuta; alto grado de interacción social.
Alto	El contacto social no siempre es primario; p. ej., profesor universitario, investigador en ciencias sociales; grado moderado de interacción social.
Moderado	Interacción social mínima, pero se necesita facilitación social; p. ej., ejecutivo de alto nivel.
Leve	Interacción social mínima y poca necesidad de preocupación acerca de los sentimientos y reacciones de los demás; p. ej., cajero en una tienda de descuento por departamentos.
Bajo	Interacción muy limitada con personas y no se requiere de roles terapéuticos o de influencia; p. ej., científico en un laboratorio o novelista.
Muy bajo	No se necesitan habilidades sociales; el ambiente laboral puede ser solitario, como programador de computadoras.

Fuente: Basado en Lowman, R. L. (1991). *The clinical practice of career assessment: Interests, abilities and personality.* Washington, DC: American Psychological Association.

- La masculinidad y la feminidad difieren significativamente entre diversos grupos ocupacionales (p. ej., Gough, 1987).

La investigación sobre la relevancia de las dimensiones de personalidad para la evaluación de carrera aún se encuentra en pañales. Sin embargo, las tendencias preliminares como aquellas que se señalan anteriormente demuestran de manera clara la relevancia de las variables de personalidad para el éxito en el trabajo. Se aconseja a los profesionales que consideren dentro de la evaluación de carrera las dimensiones de personalidad relacionadas ocupacionalmente. En resumen, la evaluación de carrera es una empresa multifacética que debe tomar en cuenta no sólo los intereses, sino también los patrones de capacidad y los rasgos de personalidad.

RESUMEN

1. El propósito de los inventarios de intereses consiste en identificar los intereses vocacionales de la persona y otros relacionados para facilitar las elecciones de carrera. Una buena adecuación entre los intereses personales y los patrones identificados de interés de una ocupación promueve el éxito en una elección profesional y la satisfacción con la misma.
2. El *Strong Interest Inventory* (SII) es la última revisión del *Strong Vocational Interest Blank* (SVIB), que apareció por primera vez en 1927. Como su predecesor, el SII utiliza codificaciones empíricas para las ocupaciones.
3. Los coeficientes de estabilidad a corto plazo de las 211 escalas ocupacionales del SII se encuentran por lo general en el rango de .90. La validez de la prueba se ha visto reforzada por la adecuación generalmente buena entre el perfil ocupacional inicial y la ocupación que en un momento dado se practica.
4. El *Jackson Vocational Interest Survey* (JVIS) utiliza un formato de reactivos de opción múltiple para reducir el impacto

de la deseabilidad social. La derivación de las 34 escalas básicas de interés fue racional y guiada por la teoría. Las escalas son razonablemente independientes entre sí y poseen coeficientes de estabilidad a corto plazo que se colocan a mitad del rango de .80. Los estudios de validez son prometedores.

5. Los 168 reactivos del *Kuder General Interest Survey* (KGIS), que se utilizan con adolescentes en bachillerato, producen 10 calificaciones de intereses generales. Los usuarios deben ser cautos en no interpretar en exceso las puntuaciones del Kuder: el coeficiente medio de estabilidad a cuatro años para las escalas es de sólo .50.
6. El VPI de 160 reactivos es un inventario objetivo, de lápiz y papel, acerca de intereses y personalidad que evalúa 11 dimensiones, incluyendo los seis temas de personalidad-ambiente denominados Realista, Investigador, Artístico, Social, Emprendedor y Convencional (RIASEC).
7. La confiabilidad del VPI es muy buena, con confiabilidades *test-retest* por escala que se encuentran en gran medida en el rango de .90. La validez de la prueba recibe apoyo de una conclusión clave en cuanto a que los individuos tienden a dirigirse a ambientes que son congruentes con sus tipos de personalidad; por ejemplo, los estudiantes tienden a seleccionar e ingresar a especialidades universitarias que coinciden con sus claves en el VPI.
8. La Búsqueda Independiente es una prueba de intereses vocacionales que aplica e interpreta el propio individuo evaluado; se basa en el modelo RIASEC; cada tema de este modelo caracteriza no sólo a un tipo de persona sino también al tipo de ambiente de trabajo que dicha persona encuentra más compatible.
9. Una medida más novedosa de autoinforme de las habilidades e intereses es la CISS. La prueba consiste de 200 reactivos de intereses y 120 reactivos de habilidades que se califican en una escala de seis puntos. La prueba produce puntuaciones *T* en numerosas escalas que incluyen las siete Escalas de orientación: Influencia, Organización, Ayuda, Creación, Análisis, Producción y Aventura.
10. La CISS también proporciona medidas independientes de intereses y habilidades para 29 Escalas básicas y 58 Escalas ocupacionales. La confiabilidad de las escalas del CISS es excepcionalmente fuerte. Los datos iniciales sobre validez que presentan los autores de la prueba son muy prometedores, como por ejemplo el hecho de que las escalas sí discriminan entre ocupaciones de una manera eficaz y significativa.
11. Los valores laborales se refieren a las necesidades, motivos y valores que influyen en la elección vocacional, la satisfacción con el trabajo y el desarrollo profesional. Una adecuación entre valores laborales y elección de carrera es importante para la satisfacción con el trabajo.
12. Una medida útil de los valores laborales es el *Minnesota Importance Questionnaire* (MIQ). Una versión de la prueba consiste de reactivos de pares de comparaciones (p. ej., Me daría una sensación de logro contra Me permitiría tomar mis propias decisiones), los cuales evalúan 20 necesidades organizadas en seis valores subyacentes relacionados con la satisfacción con el trabajo.
13. El WVI de Super cuenta con 45 reactivos y es un instrumento corto diseñado para medir 15 valores laborales en estudiantes de secundaria y preparatoria. La confiabilidad y validez del instrumento son buenas. Como en el MIQ, el WVI se interpreta en relación con los patrones de reforzamiento ocupacional.
14. La Escala de valores (VS) es un instrumento derivado de manera transcultural que se diseñó para evaluar 21 valores relacionados con los roles laborales y vitales. La prueba consiste de cinco reactivos por valor, los cuales se califican en una escala de cuatro puntos. Dado que es principalmente un instrumento para investigación, las aplicacio-

nes para orientación individual deben abordarse con precaución.

15. Para algunas personas, la elección de carrera no es una decisión separada, sino un proceso continuo. Super establece cinco etapas del desarrollo profesional: crecimiento, exploración, establecimiento, mantenimiento y declinación. En las etapas de crecimiento y exploración, los instrumentos especializados son útiles para la evaluación vocacional eficaz.

16. Lowman ha propuesto un modelo integrador de la evaluación de carrera que incluye la consideración simultánea de las áreas de intereses, capacidades y personalidad. Los patrones de capacidad, incluyendo la inteligencia social, pueden ser muy importantes en algunas vocaciones.

TÉRMINOS Y CONCEPTOS CLAVE

Modelo RIASEC p. 535
Valores laborales p. 544
Patrones de reforzamiento ocupacional p. 545
Modelo integrador p. 549
Inteligencia social p. 550

TEMA 12B
Valores en el desarrollo moral, espiritual y religioso

Evaluación de los valores vitales
Evaluación del juicio moral
Evaluación de conceptos espirituales y religiosos
Resumen
Términos y conceptos clave

En los medios populares de información se encuentran referencias frecuentes a los valores y a los cambios en dichos valores a nivel individual y nacional. Los políticos condenan la decadencia de los valores familiares, los editores de revistas denuncian la ausencia del altruismo voluntario y los columnistas censuran el resurgimiento del materialismo y la ambición. Los líderes religiosos también saltan a la palestra. Como antídoto para el cinismo global, demandan un retorno de los valores espirituales que afirman el significado de la vida. Casi todo el mundo tiene una opinión acerca de los valores —en especial con respecto a aquéllos de otras personas o grupos.

Pero ¿qué son los valores y cómo se les puede medir? Aunque existe una gran cantidad de literatura acerca de la naturaleza y definición de los valores, sorprende la poca investigación empírica sobre su medición. En general, los psicólogos definen al **valor** como una creencia compartida y duradera acerca de los modos ideales de conducta o estados meta de la existencia (Rokeach, 1980). Los valores persuaden a la acción, moldean actitudes y guían los esfuerzos para influir a los demás. Los valores

también surgen en respuesta a condiciones sociales y, por tanto, tienen cierto grado de maleabilidad (Ball-Rokeach, Rokeach y Grube, 1984).

En este tema se analizarán los aspectos clave y las pruebas importantes que se relacionan con la evaluación de los valores personales, ampliamente definidos. Se comienza con una crítica de los instrumentos generales que evalúan valores vitales —los fines o metas sociales que se considera deseable alcanzar. Después el capítulo reseña los enfoques de evaluación en los dominios moral, espiritual y religioso. Esto incluye una amplia cobertura del método clásico de Kohlberg (1981, 1984) para la medición del razonamiento moral. Se cierra con una breve revisión de la literatura no explorada acerca de la medición de conceptos espirituales y religiosos.

EVALUACIÓN DE LOS VALORES VITALES

Los valores son importantes porque proporcionan un esquema generalizado para los actos y juicios personales. Cuando se conocen los valores vitales de un individuo, pueden predecirse conductas típicas y conjeturar las probables actitudes. En un trabajo clásico sobre el tema, Rokeach (1968) subraya la importancia de los valores:

> Decir que una persona "tiene un valor" es igual a decir que tiene una creencia persistente de que un modo específico de conducta o estado meta de la existencia es preferible en términos personales y sociales a las alternativas en cuanto a modos de conducta o estados meta de la existencia. Una vez que un valor se ha internalizado, se vuelve, a nivel consciente o inconsciente, una norma o criterio para guiar los actos, para desarrollar o conservar actitudes hacia objetos y situaciones relacionadas, para justificar los actos y actitudes propios y de otras personas, para emitir un juicio moral sobre uno mismo o los demás, y para compararse uno mismo con otros. Por último, un valor es una norma que se emplea para influir en los valores, actitudes y actos de, cuando menos, algunas otras personas —por ejemplo, nuestros hijos (pp. 159-160).

Kluckhohn (1951) y Smith (1963) también han propuesto este punto de vista acerca de que los valores son, en cierto sentido, primarios y formativos.

Es más fácil definir los valores que medirlos. Unas cuantas escalas de valores han soportado la prueba del tiempo. En este capítulo se revisan tres instrumentos: el Estudio de valores es una prueba interesante que tiene principalmente importancia histórica; el *Rokeach Value Survey* (Encuesta de Valores de Rokeach) es una herramienta muy respetada para investigación; el Inventario de valores proporciona una ilustración que insta a la cautela acerca de cómo, en ocasiones, las malas pruebas logran llegar a publicación.

ESTUDIO DE VALORES

Los psicólogos se han interesado en la evaluación de los valores personales desde inicios de este siglo. No obstante, ha sido sólo en los últimos 30 años que se han desarrollado medidas de autoinforme de los valores que son sólidas en términos psicométricos. Uno de los primeros instrumentos de este estilo fue el Estudio de valores (SOV), un inventario diseñado para medir seis actitudes básicas de valoración: Teóricas (T), Económicas (E), Estéticas (Et), Sociales (S), Políticas (P) y Religiosas (R) (Allport y Vernon, 1931; Allport, Vernon y Lindzey, 1960). Estos seis valores se basaron directamente en los *Types of Men* de Spranger (1928). En este influyente libro, el intelectual alemán Eduard Spranger afirmaba que la mayoría de las personas presentan uno de los siguientes valores como el más dominante que define su personalidad:

- **Teórico (T):** el interés dominante de la persona teórica es el descubrimiento de la verdad.
- **Económico (E):** la persona económica se interesa de manera primordial en aquello que es útil.
- **Estético (Et):** la persona estética le da el mayor valor a la forma y armonía.
- **Social (S):** el amor por la gente es el valor más elevado para la persona social.

- **Político (P):** la persona política se interesa principalmente por el poder.
- **Religioso (R):** la persona religiosa le asigna el mayor valor a la unidad mística con el universo.

La escala SOV consiste de 30 preguntas que contraponen un valor con otro, además de otras 15 preguntas que requieren de la clasificación de los valores. Ejemplos de las preguntas incluyen los siguientes:

- Cuando visita un templo, ¿se siente más impresionado por una sensación generalizada de reverencia y veneración o por las características arquitectónicas y los emplomados? (Religioso contra Estético).
- En su opinión, ¿se ha avanzado más en el progreso general debido a la liberación de los esclavos, junto con el incremento en el valor que se le da a la vida individual, o debido al descubrimiento de la maquinaria de vapor, con la consecuente industrialización y rivalidad económica de los países de Europa y América? (Social contra Económico).

A partir de las respuestas a las preguntas de opción forzosa y de la clasificación de los valores, se traza un perfil de valores de manera ipsativa, representando las fortalezas relativas de los seis valores para cada individuo.

Lubinski, Schmidt y Benbow (1996) demostraron el mérito del examen de los valores con el SOV en un estudio de seguimiento a 20 años con 203 adolescentes intelectualmente talentosos. Se evaluó inicialmente a la muestra de individuos dotados cuando tenían 13 años de edad y después de nuevo en la adultez, a la edad de 33 años. En general, los seis temas revelaron estabilidad significativa a lo largo de este periodo, con correlaciones medias interindividuales de .37 para diversos temas. Esto es notable, dado que se supone que los años de la adolescencia y adultez joven son un periodo de confusión y cambio, en especial en cuanto a valores personales, a medida que las personas jóvenes se esfuerzan por encontrar una identidad. Las diferencias por sexo fueron notables: Los varones tendieron a virar hacia un perfil T-E-P en la adultez, mientras que las mujeres tendieron a virar a un perfil Et-S-R. Aun así, se observó un patrón común en todos los participantes, en el que los valores Estéticos y Económicos adquirieron mayor prominencia en la adultez joven y los valores Políticos y Sociales revelaron menor dominio.

El Estudio de valores ha provocado considerable discusión como una herramienta de demostración en el salón de clases dentro de cursos de psicología, pero en otros sentidos no ha sido una prueba importante. Uno de los principales problemas con el instrumento es que los seis valores se definen de manera vaga y demasiado general como para tener un uso práctico. No obstante, la prueba inspiró a otros autores a desarrollar enfoques más complejos y amplios de la evaluación de los valores. Uno de aquellos que reconocieron tener una deuda con Allport y el Estudio de Valores fue Milton Rokeach.

ENCUESTA DE VALORES DE ROKEACH

Rokeach (1973) definió dos tipos de valores, instrumentales y terminales. Los primeros son modos deseables de conducta, mientras los segundos son estados meta deseables de la existencia. Por ejemplo, la ambición es un valor instrumental, y la seguridad familiar es un valor terminal. Para el diseño de la Rokeach Value Survey (Encuesta de Valores de Rokeach), se obtuvo una lista final de 18 valores instrumentales al condensar 555 nombres de "rasgos de personalidad" en sinónimos aproximados. La lista final de 18 valores terminales se obtuvo de un examen de la literatura y de otros enfoques subjetivos, impresionistas. Los 36 valores se listan en el cuadro 12–6.

Aunque los valores individuales no se definen en detalle, cada uno se acompaña de una frase corta o de sinónimos para aclarar el reactivo a las personas evaluadas. Por ejemplo, el primero de los valores terminales se presenta como sigue: "UNA VIDA CÓMODA" (una vida próspera), Llenar el estudio es sumamente

Cuadro 12–6. Los 36 constructos de valor de la Encuesta de Valores de Rokeach, Forma D

Valores terminales	
Una vida cómoda	Armonía interna
Una vida emocionante	Amor maduro
Sensación de logro	Seguridad nacional
Un mundo en paz	Placer
Un mundo de belleza	Salvación
Igualdad	Respeto hacia uno mismo
Seguridad familiar	Reconocimiento social
Libertad	Verdadera amistad
Felicidad	Sabiduría
Valores instrumentales	
Ambicioso	Imaginativo
Tolerante	Independiente
Capaz	Intelectual
Alegre	Lógico
Limpio	Amante
Valiente	Obediente
Indulgente	Cortés
Servicial	Responsable
Honrado	Autocontrolado

sencillo. A las personas que responden la prueba se les pide que clasifiquen por separado los 18 valores terminales y 18 valores instrumentales con base en "su importancia para usted, como principios guía en su vida". Los valores están impresos en etiquetas adhesivas (para la Forma D). Los individuos simplemente desprenden las etiquetas y las clasifican en orden de importancia, retirándolas y volviéndolas a colocar según sea necesario. La categoría para cada reactivo se vuelve la calificación para ese valor. No se permiten empates, de modo que las calificaciones de valor irán de 1 a 18, donde las puntuaciones más bajas indican mayor importancia.

La confiabilidad del Estudio de valores puede enfocarse de dos maneras. La primera es la estabilidad temporal de las clasificaciones para las personas individuales. Para este enfoque, la escala se aplica en dos ocasiones y se correlacionan los dos conjuntos de clasificaciones para cada individuo. Al emplear este enfoque con cuatro grupos de estudiantes universitarios (intervalos de repetición de prueba de tres semanas a cuatro meses), Rokeach (1973) informó correlaciones medianas *test-retest* que iban de .76 a .80 para los valores terminales, y de .65 a .72 para los valores instrumentales. La segunda manera de examinar la confiabilidad consiste en calcular por separado la confiabilidad *test-retest* de las calificaciones individuales de valor, entre todas las personas evaluadas. Con este enfoque, la confiabilidad de las escalas individuales es menor, aproximadamente de .65 para los valores terminales y de .56 para los valores instrumentales (Rokeach, 1973). Estas confiabilidades son bastante bajas en comparación con instrumentos que tienen más reactivos por escala —lo cual no es sorprendente. Después de todo, las "escalas" del Estudio de valores consisten cada una de un solo reactivo. Sin embargo, con confiabilidades así de bajas, el Estudio de valores debería utilizarse sólo con propósitos de investigación como la descripción o comparación de valores grupales. La interpretación con propósitos de orientación no tiene base.

En un fascinante ejemplo de su aplicación en investigación, Rokeach y sus colaboradores utilizaron el Estudio de valores para medir los efectos de ver un solo programa de televisión con duración de 30 minutos sobre los valores, actitudes y conductas (Ball-Rokeach, Rokeach y Grube, 1984). El programa de televisión, presentado por Ed Asner y que se llamó "*The Great American Values Test*", se diseñó de manera particular para influir en las calificaciones de los espectadores acerca de la importancia de valores terminales como libertad e igualdad. Por ejemplo, acerca de una presentación gráfica en pantalla, la cual indicaba que los estadounidenses habían clasificado a la libertad en tercer lugar y a la igualdad en doceavo, en promedio, entre los 18 valores terminales, Asner comentó:

> Los estadounidenses sienten que la libertad es muy importante. La clasifican en tercer lugar. Pero también sienten que la igualdad es considera-

blemente menos importante ... la clasifican en doceavo lugar. Dado que la mayoría de los estadounidenses valoran mucho más a la libertad que a la igualdad, la pregunta sería: ¿qué quiere decir eso? ¿Esto sugiere que los estadounidenses en general están mucho más interesados en su propia libertad que en la de otras personas? ¿Existe una contradicción dentro del pueblo de Estados Unidos entre su amor a la libertad y su menor amor hacia la igualdad?

Al comparar sus valores con estos resultados, usted debería poder decidir por sí mismo si está de acuerdo con los sentimientos del estadounidense promedio acerca de la libertad y la igualdad (Ball-Rokeach y cols., 1984).

Un análisis completo de este estudio implicaría una extensa desviación con respecto al tema de las pruebas psicológicas. No obstante, es posible que el lector aprecie un rápido resumen. Los autores utilizaron un diseño de preprueba y posprueba sumamente controlado con ciudades experimentales y control para determinar los efectos de ver el programa. Para los espectadores que vieron el programa sin interrupción, las clasificaciones medias para igualdad fueron de 11.0 a 9.3, mientras que en el caso de los no espectadores, las clasificaciones para este valor fueron bastante estables. Varias otras verificaciones experimentales (p. ej., solicitar donaciones para proporcionar oportunidades culturales a niños afroestadounidenses) también confirmaron un cambio real en valores. Este estudio es un buen ejemplo del tipo de investigación social para el que el Estudio de valores es muy adecuado.

LIMITACIONES DE LA ENCUESTA DE VALORES DE ROKEACH

Ya se ha mencionado que las escalas individuales del Estudio de valores poseen confiabilidad marginal —lo cual significa que el instrumento no debería utilizarse para orientación individual. Varias otras limitaciones se derivan de la naturaleza ipsativa de la prueba. El lector recordará que una prueba ipsativa es aquélla en la que el promedio de las escalas siempre es el mismo para todas las personas evaluadas. En particular, el rango promedio para los 18 valores instrumentales siempre será de 9.5 y lo mismo ocurre con los valores terminales. Por definición, cuando un individuo evaluado le da mayor clasificación a algunas escalas, otras deben recibir una clasificación baja. En este proceso se pierde cualquier medida absoluta del valor para ese individuo. Supóngase, por ejemplo, que se puede medir la fortaleza absoluta de los 18 valores instrumentales en una escala de 1 a 100 (nota: esto no es posible con el Estudio de valores). Considérese el caso donde un individuo *A* tiene una fortaleza absoluta de 99 para ambicioso y de 98 para obediente, con todos los demás valores por debajo de 90; mientras que un individuo B tiene una fortaleza absoluta de 39 para ambicioso y de 19 para obediente, con todos los demás valores por debajo de 10. Casi con toda seguridad, el individuo *A* valoraría la ambición y la obediencia en grado elevado, mientras que el individuo *B* le da un valor modesto a la ambición y devalúa la obediencia. De hecho, al individuo *B* se le podría caracterizar como prácticamente carente de valores. No obstante, ambas personas recibirían valores de 1 para ambicioso y 2 para obediente. El Estudio de valores no es sensible a las diferencias de magnitud dentro de personas individuales y tampoco capta diferencias de grado entre individuos.

Braithwaite y Law (1985) llaman la atención hacia debilidades adicionales del Estudio de valores. Señalan que el inventario omite varios valores importantes, incluyendo bienestar físico, derechos individuales, frugalidad y despreocupación. Quizás es más importante su crítica acerca de que la prueba Rokeach depende de un solo reactivo para cada valor, en vez de utilizar índices con reactivos múltiples para los constructos de valor. Estos autores proponen un instrumento alternativo (basado en el enfoque de Rokeach) que supuestamente engloba cualidades psicométricas mejoradas para la medición de los valores personales.

INVENTARIO DE VALORES

El Inventario de valores se diseñó para evaluar seis valores en gerentes profesionales: teórico,

de poder, de logro, humano, industrial y financiero. El instrumento consiste de 28 conjuntos de tres citas semejantes a las siguientes:

- La menor brizna de verdad representa el amargo esfuerzo y desventura de algún hombre.
- Nunca hacemos nada bien si pensamos en la manera de hacerlo.
- Todo trabajo es noble y sólo el trabajo es noble.

La persona evaluada indica su grado de acuerdo con cada afirmación. Se distribuyen tres puntos entre cada conjunto de tres citas: un punto para cada una si todas se refrendan por igual, tres puntos para una de ellas si es la única que se refrenda y así sucesivamente. Como podrá haber adivinado el lector, cada afirmación representa un valor específico. Al sumar el número de puntos que se conceden a cada uno de los seis valores entre los 28 conjuntos de reactivos, se obtiene un perfil de calificaciones de valor.

Los revisores se han mostrado mordaces en sus críticas para el Inventario de valores, señalando que la prueba carece de una fundamentación teórica y ni siquiera posee validez nominal, no se diga validez técnica (Muchinsky, 1985; Tziner, 1985). No se menciona al instrumento como algún tipo de actitud ruin, sino simplemente para ilustrar que, en ocasiones, las pruebas cuestionables llegan a publicarse. Los usuarios de las pruebas deben ejercitar su juicio antes de adoptar nuevos instrumentos.

EVALUACIÓN DEL JUICIO MORAL

ESCALA DE JUICIO MORAL

Kohlberg ha propuesto una de las pocas teorías del desarrollo moral que es completa y se basa en un fundamento empírico (Colby, Kohlberg, Gibbs y Lieberman, 1983; Kohlberg, 1958, 1981, 1984; Kohlberg y Kramer, 1969). Aunque se ocupó más de los problemas del desarrollo moral basados en la teoría que de los matices de la medición estandarizada, Kohlberg generó un método de evaluación que se utiliza de manera amplia y que se ha discutido intensamente. Se revisará la fundamentación subyacente para su herramienta de medición y también se analizarán las propiedades psicométricas del instrumento. Además, se examinará de manera breve una adaptación más objetiva del enfoque de Kohlberg, la cual se conoce como *Defining Issues Test* (Prueba de cuestiones determinantes) (Rest, 1979; Rest y Thoma, 1985).

ETAPAS DEL DESARROLLO MORAL

La teoría de Kohlberg partió de la teoría de Piaget (1932) acerca de los estadios del desarrollo moral en la infancia. Este autor extendió las etapas hasta la adolescencia y adultez. A fin de explorar el razonamiento sobre problemas morales difíciles, Kohlberg diseñó una serie de **dilemas morales**. Uno de los más famosos es el de Heinz y el farmacéutico:

> En Europa, una mujer se encontraba cerca de la muerte debido a un tipo especial de cáncer. Había un medicamento que los doctores consideraban que podría salvarla. Era una forma de radio que un farmacéutico en el mismo pueblo había descubierto recientemente. Era costoso producir el medicamento, pero el farmacéutico cobraba 10 veces más de lo que le costaba mandar hacer la medicina. Pagaba $200 por el radio y cobraba $2 000 por una dosis pequeña de la droga. El esposo de la mujer enferma, Heinz, acudió con todas las personas a las que conocía para pedir prestado el dinero, pero sólo pudo reunir cerca de $1 000, que es la mitad del costo. Le dijo al farmacéutico que su mujer estaba muriendo y le pidió que le vendiera la medicina más barata o que le permitiera pagarle después. Pero el farmacéutico dijo, "No, yo descubrí la medicina y voy a obtener dinero de ella". Por ese motivo, Heinz se sintió desesperado e irrumpió en la tienda del farmacéutico para robar el medicamento para su esposa (Kohlberg y Elfenbein, 1975).

Después de leer o escuchar esta historia, se hacen una serie de preguntas de sondeo a la persona evaluada. Éstas podrían ser como las siguientes: ¿Heinz debería haber robado la medicina? ¿Qué sucede si Heinz no amaba a su esposa? ¿Eso cambiaría la situación? ¿Qué pasa

si la persona que estaba muriendo fuera una desconocida? ¿Heinz debería robar de todas maneras el medicamento? Con base en las respuestas a este y otros dilemas, Kohlberg concluyó que existen tres niveles principales de razonamiento moral, con dos subetapas dentro de cada nivel (cuadro 12–7). Un empleo para este instrumento de medición, la *Moral Judgement Scale* (Escala de Juicio Moral), consiste en determinar la etapa de razonamiento moral de la persona evaluada.*

La Escala de Juicio Moral consiste de varios dilemas hipotéticos como el de Heinz y el farmacéutico, los cuales se presentan uno a la vez (Colby, Kohlberg, Gibbs y colaboradores, 1978). En su última revisión, la Escala se presenta en tres versiones llamadas Formas A, B y C. La calificación es bastante compleja y se basa en el juicio del examinador acerca de las respuestas en relación con extensos criterios descritos en un detallado manual de calificación (Colby y Kohlberg, 1987). Aunque existen varias dimensiones diferentes para la calificación, el elemento que se cita con más frecuencia en los estudios de investigación es la etapa general de razonamiento moral que caracteriza a la persona evaluada.

CRÍTICA DE LA ESCALA DE JUICIO MORAL

Las primeras versiones de la Escala de juicio moral sufrieron graves defectos de calificación e interpretación. Por ejemplo, en su tesis doctoral, Kohlberg (1958) propuso dos sistemas de calificación: uno que utilizaba una oración o pensamiento completo como la unidad de calificación y el otro que dependía de la calificación global de todas las expresiones del individuo como la unidad de análisis. Ninguno de los enfoques fue completamente satisfactorio y las primeras revisiones de la escala tuvieron críticas justificables acerca de la confiabilidad y validez del instrumento (Kurtines y Greif, 1974).

En respuesta a estas críticas, Kohlberg y sus colaboradores desarrollaron un sistema de calificación que no tiene paralelo en cuanto a su claridad, detalle y complejidad (Rest, 1986). Por fortuna, dado que los dilemas de la Escala de Juicio Moral han permanecido constantes a lo largo de los años, es posible aplicar los nuevos sistemas de calificación a los viejos datos. La capacidad de volver a analizar los viejos datos y compararlos con los nuevos es inestimable para determinar la confiabilidad y validez de una escala existente. Uno de los estudios más importantes en este sentido fue el publicado por Kohlberg y colaboradores (Colby y cols., 1983).

Esta investigación presenta los resultados de utilizar el nuevo sistema de calificación en un estudio longitudinal que abarcó más de 20 años. Éstos fueron impresionantes y ofrecen

Cuadro 12–7. Niveles y etapas del desarrollo moral según Kohlberg

Nivel 1: Preconvencional

Etapa 1. Orientación hacia el castigo y la obediencia: las consecuencias físicas determinan lo que es bueno o malo.

Etapa 2. Orientación al relativismo instrumental: es bueno aquello que satisface las propias necesidades.

Nivel 2: Convencional

Etapa 3. Orientación hacia la concordancia interpersonal: es bueno aquello que complace o ayuda a los demás.

Etapa 4. Orientación hacia la "ley y el orden": Es bueno aquello que conserva el orden social y cumplir con el propio deber.

Nivel 3: Posconvencional o de principios

Etapa 5. Orientación legal-de contrato social: Los valores acordados por la sociedad determinan lo que es bueno.

Etapa 6. Orientación de principios éticos universales: aquello que es correcto es una cuestión de consciencia y se obtiene de principios universales.

Fuente: Basado en Kohlberg (1984).

* Aunque la Escala de Juicio Moral se ha utilizado ampliamente para la investigación empírica, Kohlberg (1981, 1984) sugiere que su aplicación más valiosa es para la promoción de la autocomprensión y el desarrollo del razonamiento moral en el individuo evaluado.

fuerte sustentación a la confiabilidad y validez del instrumento. Las correlaciones *test-retest* para las tres formas se colocaron en la parte superior del rango de .90, al igual que las correlaciones intercalificadores. Las calificaciones longitudinales de los individuos sometidos a prueba en intervalos de tres a cuatro años durante 20 años revelaron tendencias consistentes con la teoría. De los 58 individuos, 56 mostraron cambio en sentido progresivo y ninguno saltó alguna etapa. Lo que es más, sólo 6% de las 195 comparaciones mostraron cambios en sentido inverso entre dos sesiones de prueba. La consistencia interna de las calificaciones también fue excelente: cerca de 70% de éstas se colocaron en una etapa y sólo 2% se dispersaron más allá de dos etapas adyacentes. La calificación de la prueba alfa de Cronbach se colocó a la mitad del rango de .90 para las tres formas. Nisan y Kohlberg (1982) corroboraron estas conclusiones. Heilbrun y Georges (1990) también dan un informe favorable acerca de la validez de la Escala de Juicio Moral, en la medida en que el desarrollo posconvencional se correlaciona con niveles más elevados de autocontrol, como se predeciría a partir del hecho de que las personas moralmente maduras se oponen con frecuencia a la presión social o a las restricciones legales. En resumen, la Escala de Juicio Moral es confiable, consistente a nivel interno y posee una coherencia evolutiva confirmatoria de la teoría.

PRUEBA DE CUESTIONES DETERMINANTES

La Defining Issues Test; DIT (Prueba de Cuestiones Determinantes) es similar a la Escala de Juicio Moral, pero incorpora un formato de calificación mucho más sencillo y completamente objetivo (Rest, 1979, 1986). La persona evaluada lee una serie de dilemas morales similares a los diseñados por Kohlberg y después elige una acción apropiada para cada uno. Por ejemplo, un dilema implica a una paciente que tiene una agonía muy dolorosa debida al cáncer. En sus momentos de lucidez pide que se le aplique una sobredosis de morfina para acelerar su muerte. ¿Qué debería hacer el médico? Se listan tres opciones del siguiente tipo:

____ Debería darle una dosis mortal.
____ No debería darle una sobredosis.
____ No puede decidir.

La opción de la persona evaluada no entra directamente en la determinación de la calificación de juicio moral. El propósito real de forzar una elección consiste en hacer que la persona examinada piense acerca de la importancia de diversos factores para tomar la decisión. Después de optar por la acción apropiada, el examinando califica la importancia de varios factores en una escala Likert de cinco puntos: gran, mucha, alguna, poca o ninguna importancia. Los factores son diferentes para cada dilema. Los factores difieren en el nivel de juicio moral que implican y van desde la etapa 1 hasta la etapa 6 de Kohlberg. En el caso del dilema anterior, los factores incluyen cuestiones tales como las siguientes:

____ Si el médico puede hacerlo parecer un accidente.
____ La sociedad puede permitir que las personas terminen con su vida cuando lo deseen.
____ Si la familia de la mujer favorece darle la sobredosis o no.

Estas clasificaciones forman la base para generar varias calificaciones cuantitativas que tienen que ver con el juicio moral de la persona evaluada. La calificación de más amplia utilización es la **P**, que es el porcentaje de razonamiento de principios. El rango de confiabilidad de las calificaciones **P** varía de **P** de .71 a .82 en los estudios *test-retest* (Rest, 1979, 1986). La validez ha sido estudiada mediante contrastar grupos conocidos que se sabe son diferentes en el razonamiento de principios. Por ejemplo, se encontró que los estudiantes de posgrado en filosofía moral y en ciencias políticas, estudiantes universitarios en general, estudiantes de preparatoria y estudiantes de secundaria diferían de manera apropiada y sistemática en cuanto a la calificación **P**. En estu-

dios longitudinales, se encontraron tendencias significativas en sentido progresivo a lo largo de seis años y cuatro evaluaciones. De manera reciente, Rest ha recomendado una nueva medida del juicio moral, el índice N2, que se calcula con base en varias fórmulas complejas que utilizan tanto datos de clasificación como de calificación. Los dos índices se correlacionan de manera elevada en el rango del .90. No obstante, en un análisis retrospectivo de estudios previos, el índice N2 superó al índice **P** por un margen sustancial (Rest, Thoma, Narvaez y Bebeau, 1997).

Se han publicado más de 600 artículos sobre la Prueba de cuestiones determinantes (McCrae, 1985; Moreland, 1985; Sutton, 1992). En general, se considera que el instrumento es una alternativa a la Escala de Juicio Moral de Kohlberg, en particular para la investigación de diferencias grupales en razonamiento moral. Sin embargo, los revisores señalan varias advertencias acerca de la DIT (Sutton, 1992; Westbrook y Bane, 1992). Primero, la prueba utiliza dos dilemas de la guerra de Vietnam y, por tanto, es un poco anticuada. Muchos individuos jóvenes tienen poco conocimiento (y quizá ningún interés) en cuanto a este tema y puede resultarles difícil identificarse con estas cuestiones. Otro dilema —el caso clásico de si Heinz debiera robar el medicamento para salvar la vida de su esposa— también tiene un valor dudoso ya que se le ha dado amplia publicidad y se le ha publicado en textos universitarios. Una proporción significativa de examinandos potenciales conocen más que bien este dilema moral.

Richards y Davison (1992) han enfatizado la cuestión de que la DIT tiene un prejuicio contra individuos de convicciones religiosas conservadoras. Ciertamente se ha establecido que las personas con creencias religiosas conservadoras o fundamentalistas tienden a obtener calificaciones inferiores al promedio en la calificación **P** de la Prueba de cuestiones determinantes (Getz, 1984; Richards, 1991). Según Richards y Davison (1992), la razón para esto es que los reactivos de la Etapa 3 y 4 poseen (de manera no intencional) fuertes implicaciones teóricas que causan que los individuos fundamentalistas refrenden los reactivos, con lo cual reducen su calificación en la prueba. Considérense los reactivos que se dirigen al razonamiento de la etapa 4, relacionada con la orientación de "ley y orden", que hace una equivalencia entre lo que es "moral" y el cumplir con el propio deber y mantener el orden social. Mientras que las personas no religiosas podrían estar de acuerdo con las leyes terrenales (y refrendar los reactivos de la etapa 4) debido a que creen que las autoridades legales definen lo que es correcto y moral, las minorías religiosas como los mormones creen que respetar las leyes terrenales es una obligación teológica y religiosa que fluye directamente de los artículos de fe en su religión:

> ... aunque los mormones le dan un alto valor a obedecer la ley y apoyar a las autoridades legales, éste se debe a su creencia religiosa de que Dios les ordena hacerlo y no debido a que crean, como los individuos con un pensamiento que realmente se encuentra en la Etapa 4, que las leyes terrenales o las autoridades legales definen lo que es correcto o moral (Richards y Davison, 1992, 470).

Estos investigadores demuestran de manera empírica que ciertos reactivos de la DIT miden un constructo diferente para las personas con creencias religiosas conservadoras que para la población general. En consecuencia, la validez de la prueba para estos grupos está abierta a discusión.

Una crítica relacionada acerca de la DIT es la escasez de normas relativas a grupos minoritarios. Por último, Westbrook y Bane (1992) afirman que el manual técnico de la DIT carece de los detalles esenciales necesarios para valorar la idoneidad de la prueba. A pesar de estas críticas, la DIT es una prueba sumamente respetada, en particular para la investigación sobre razonamiento moral.

EVALUACIÓN DE CONCEPTOS ESPIRITUALES Y RELIGIOSOS

Dentro del campo de la psicología, los temas trascendentales como el bienestar espiritual o

la madurez de la fe nunca han recibido atención de los investigadores convencionales. Hace 50 años, Gordon Allport (1950) se lamentaba de que el tema de la religión "parecía haberse ocultado" entre los intelectuales e investigadores académicos:

> Sin importar cuál pueda ser la razón, la persistencia de la religión en el mundo moderno parece ser motivo de vergüenza para los eruditos de la actualidad. Incluso los psicólogos, para quienes supuestamente ninguna de las cuestiones humanas les es ajena, están en probabilidad de retraerse cuando se saca a relucir el tema (p. 1).

La situación ha mejorado poco en la época contemporánea. Por ejemplo, excepto por unas cuantas revistas especializadas, los temas espirituales y religiosos están casi ausentes en la literatura psicológica.

No obstante, los investigadores no tienen el derecho de retirarse del campo, dada su importancia para la persona promedio. Considérense las estadísticas sobre creencias religiosas en EUA, estables desde 1944, cuando se pusieron en práctica por primera vez las encuestas (Hoge, 1996):

- La creencia en Dios ha permanecido constante en aproximadamente 95% de la población.
- De 75 al 77% de los adultos se adhieren a la creencia en la divinidad de Jesucristo.
- La creencia en la vida después de la muerte persiste en cerca de 75% de la población.

No se encuentran disponibles estadísticas comparables a nivel mundial, pero parece probable que el porcentaje de individuos creyentes (ya sean musulmanes, budistas, hindúes, judíos u otros) sea muy alto. La mayoría de las personas se adhieren a alguna perspectiva religiosa en la vida y, con toda seguridad, esto debe tener alguna relación con su adaptación, conducta y perspectiva.

Por desgracia, el campo de la psicología, incluyendo el área de especialidad de las pruebas, ha conservado en gran medida una actitud indiferente hacia este importante aspecto de la experiencia humana. Peor aún, en muchos círculos intelectuales, la aceptación de sentimientos espirituales o religiosos se considera como evidencia de psicopatología. Entre otros, Sigmund Freud se adhirió a una perspectiva cínica de la religión en su ensayo adecuadamente titulado *The Future of an Illusion* (**El porvenir de una ilusión**; 1927/1961). Sin embargo, para muchas personas la conexión con lo trascendente es esencial para el significado de la vida. Esto ocurre especialmente en tiempos de cautiverio, como cuando la aniquilación de uno mismo toca a la puerta. Considérese la experiencia de Viktor Frankl (1963), un superviviente de un campo de exterminio nazi y figura fundadora de la psicología existencial. En algún momento de la segunda Guerra Mundial tuvo que resignarse a perder su abrigo con un apreciado manuscrito en los bolsillos, a cambio de los andrajos hechos pedazos de un interno enviado a la cámara de gas:

> En lugar de las muchas páginas de mi manuscrito, encontré en un bolsillo del recientemente adquirido abrigo una sola página arrancada de un libro hebreo de oraciones que contenía la principal oración Judía, *Shema Yisrael*. ¿Cómo debería haber interpretado tal "coincidencia", si no como un reto a vivir mis pensamientos en vez de sólo ponerlos en papel?

En lo que resta de este tema, se toma la perspectiva de que las dimensiones espirituales y religiosas de la vida satisfacen a menudo propósitos constructivos y que la evaluación dentro de estas áreas es digna de estudio adicional.

FUNDAMENTACIÓN PARA LA EVALUACIÓN RELIGIOSA Y ESPIRITUAL

Los investigadores académicos justifican la evaluación de las dimensiones religiosa y espiritual como un medio para buscar la verdad acerca de la conducta humana. Su principal motivo es la curiosidad intelectual y su meta es la comprensión del papel de la religión y de la espiritualidad en los asuntos humanos. Pero, ¿existe algún tipo de razón práctica para intentar medir las dimensiones religiosa o espiritual en el **individuo**? Por ejemplo, ¿los profesionales clínicos, como psicólogos o trabajadores

sociales, obtienen alguna información útil de la evaluación de los antecedentes religiosos y espirituales de sus clientes?

Richards y Bergin (1997) ofrecen varios argumentos convincentes a favor de que los terapeutas evalúen los antecedentes religiosos y espirituales de sus clientes. En términos específicos, esta forma de evaluación serviría para:

1. Aumentar la empatía al ayudar a los terapeutas a comprender mejor la perspectiva global de sus clientes.
2. Identificar y evaluar el impacto de las orientaciones religiosas-espirituales sanas y patológicas de sus clientes.
3. Determinar si las creencias religiosas y espirituales y la comunidad pueden proporcionar apoyo a los clientes.
4. Identificar las posibles intervenciones espirituales que pueden emplearse en terapia para ayudar a los clientes.
5. Determinar si éstos poseen dudas o preocupaciones espirituales no resueltas que necesiten atenderse.

En general, Richards y Bergin (1997) afirman que los clínicos necesitan atender a la persona completa, a fin de proporcionar la mejor respuesta posible a los problemas emocionales y psicológicos. Debido a que la mayoría de los clientes llevan sus preocupaciones religiosas y espirituales a la sesión de terapia, los clínicos que evalúan estas dimensiones de la existencia humana pueden estar mejor preparados para proporcionar un servicio eficaz.

PERSPECTIVA HISTÓRICA GENERAL ACERCA DE LA EVALUACIÓN RELIGIOSA

El interés en la psicología de la religión puede rastrearse a inicios del siglo XX cuando William James (1902) escribió su obra maestra, *The Varieties of Religious Experience.* En este libro, James catalogó las muchas maneras en que los seres humanos revelan su interés por cuestiones trascendentales. Su conclusión general fue que la religión es "un órgano esencial de nuestra vida, que lleva a cabo una función que ninguna otra parte de nuestra naturaleza puede satisfacer con tanto éxito".

Aunque muchos autores han ofrecido análisis psicológicos de la religión desde los escritos originales de James, no fue sino hasta el decenio de 1960 que comenzaron a aparecer escalas para la evaluación de las variables religiosas (Wulff, 1996). Una de las primeras medidas de este tipo fueron las *Allport-Ross Religious Orientation Scales* (Escalas Allport-Ross de orientación religiosa), que proponían evaluar dos dimensiones de la expresión religiosa, la **intrínseca** y la **extrínseca** (Allport y Ross, 1967). Se consideraba que las personas con religiosidad intrínseca **vivían** su religión (p. ej., para encontrar significado, dirección, perspectiva), mientras que se creía que aquellas con religiosidad extrínseca **utilizaban** su religión (p. ej., para buscar seguridad, estatus, sociabilidad). En sus primeros escritos sobre el tema, Allport hizo referencia a la forma intrínseca como una orientación religiosa genuina o madura, mientras que la extrínseca se consideraba inmadura. Posteriormente hizo a un lado las denominaciones madura-inmadura, porque las etiquetas parecían demasiado críticas.

El ímpetu para el desarrollo de estas escalas fue la angustiosa observación de Allport de una relación positiva entre religiosidad (en ciertas formas) y actitudes de autoritarismo, intolerancia y prejuicio. Como una persona devotamente religiosa, Allport se mostraba convencido de que sería poco común que los individuos con una orientación religiosa intrínseca albergaran dichas actitudes. Después de todo, un precepto esencial de casi toda creencia religiosa es una actitud de amor hacia el prójimo. En la fe cristiana, esta perspectiva se resume en la famosa sentencia: "Amarás a tu prójimo como a ti mismo" (Marcos 12:31). No obstante, la evidencia abrumadora para Allport fue que algunos individuos religiosos revelaban odio, intolerancia y prejuicio hacia sus semejantes. Los objetivos comunes de estas actitudes maliciosas eran las minorías raciales, los judíos y las personas homosexuales, entre otros. Allport

razonó que las personas religiosas con actitudes intolerantes poseían predominantemente una orientación religiosa extrínseca; es decir, su fe satisfacía metas externas como la posición dentro de la comunidad, la pertenencia a un grupo y similares. La investigación sobre esta hipótesis (de que las personas con religiosidad extrínseca mostrarían mayor autoritarismo, intolerancia y prejuicio que aquéllas con religiosidad intrínseca) requirió de las herramientas apropiadas. Para tal propósito, Allport y sus colaboradores desarrollaron las escalas de Orientación religiosa.

Ejemplos de los tipos de reactivos en la escala Extrínseca de 11 de ellos y en la escala Intrínseca de 9 son los siguientes:

- La iglesia es importante como lugar para desarrollar buenas relaciones sociales (extrínseca).
- En ocasiones me parece necesario comprometer mis creencias religiosas por razones económicas (extrínseca).
- Intento firmemente llevar mi religión a otros aspectos de mi vida (intrínseca).
- Mi religión es importante porque proporciona significado para mi vida (intrínseca).

Aunque en un principio se le diseñó en un formato de sí-no, las aplicaciones modernas de estas escalas utilizan un continuo de nueve puntos desde (1) fuertemente en desacuerdo hasta (9) fuertemente de acuerdo (Batson, Schoenrade y Ventis, 1993).

La investigación sobre las escalas de Orientación religiosa no ha proporcionado un fuerte apoyo a la hipótesis original de Allport (Wulff, 1996). De hecho, varios estudios han mostrado que las personas con puntuaciones más altas en la escala Intrínseca revelan, de hecho, **mayores** niveles de autoritarismo, intransigencia y prejuicio hacia afroestadounidenses, homosexuales y lesbianas. Hunsberger (1995) concluye que no es la religión en sí la que da lugar al prejuicio, como tampoco la orientación religiosa intrínseca/extrínseca. En vez de ello, "es la manera en la que se sustentan las creencias religiosas lo que parece asociarse de manera más directa con el prejuicio y esto se explica mejor por medio del vínculo estrecho con la tendencia hacia el fundamentalismo y al autoritarismo de derecha". En términos específicos, este autor vincula el prejuicio hacia las minorías con las tradiciones religiosas autoritarias que promueven una verdad absoluta, dividen al mundo en fuerzas del "Bien" y del "Mal" y evitan la complejidad y la duda en sus sistemas de creencias. Estos aspectos de la expresión religiosa no se miden por lo común a través de pruebas de lápiz y papel.

En un estudio reciente, Genia (1993a) concluyó que las escalas combinadas de Orientación religiosa (que consisten de 20 reactivos) miden **tres** factores de expresión religiosa y no los dos propuestos por Allport (Intrínseca y Extrínseca). Esta investigadora llevó a cabo un análisis factorial de los resultados de la prueba de 309 personas entre 17 y 83 años (edad media de 29 años), provenientes de diversas tradiciones religiosas. Los nueve reactivos en la escala Intrínseca original se agruparon de manera muy adecuada, lo cual da lugar a un factor robusto de orientación religiosa intrínseca. Pero los 11 reactivos en la escala Extrínseca original se separaron en dos subfactores independientes que Genia (1993a) denominó uso de la religión para beneficios personales (Ep) y para recompensas sociales (Es). La autora recomendó transformar el instrumento original en una prueba de tres factores (Intrínseca, Extrínseca-personal, Extrínseca-social) al añadir y eliminar unos cuantos reactivos. No obstante, la mayoría de las autoridades en el campo han concluido que las escalas de Allport satisficieron una valiosa función al identificar dimensiones clave de la experiencia religiosa y estimular la investigación, pero en la actualidad ya no son útiles.

RELIGIÓN COMO BÚSQUEDA

Cada vez es más frecuente que la base conceptual para la distinción entre orientación religiosa Intrínseca y Extrínseca se ponga en tela de juicio. Kirkpatrick y Hood (1990) resumen las

principales críticas teóricas y metodológicas para las escalas de la siguiente manera:

- Una falta de claridad conceptual en cuanto a aquello que se supone que las escalas Intrínseca y Extrínseca miden. ¿Son tipos de motivación (es decir, los motivos asociados con la creencia y práctica religiosa) o variables de la personalidad (es decir, aspectos generales de la conducta o participación institucional) o alguna otra cosa?
- Una confusión acerca de la relación entre las escalas Intrínseca-Extrínseca. En particular, ¿son extremos opuestos de una sola dimensión bipolar o las escalas miden dimensiones independientes (de modo que sería concebible la posibilidad de que algunas personas pudieran tener altas calificaciones en **ambas**)?

Otros problemas mencionados incluyen la debilidad en la estructura factorial, confiabilidad y validez de constructo de las escalas; la dependencia excesiva en una dicotomía entre "religión buena" y "religión mala" y la insensatez de definir y estudiar la religiosidad independientemente del contenido de la creencia (Kirkpatrick y Hood, 1990).

En respuesta a las limitaciones de las escalas de Orientación religiosa, Batson y colaboradores (1993) desarrollaron una medida de una tercera orientación religiosa conocida como Búsqueda. Estos investigadores consideran a la Búsqueda como una perspectiva religiosa más madura y flexible que las orientaciones Intrínseca y Extrínseca. De hecho, Allport reconoció los elementos inherentes en esta orientación, pero no los incorporó en su escala Intrínseca. La **Búsqueda** se caracteriza por la complejidad, duda y vacilación como maneras de ser religioso. Esta orientación sugiere:

> ... un enfoque que implica enfrentar de manera honesta las preguntas existenciales en toda su complejidad, mientras que al mismo tiempo se resisten las respuestas irrefutables y convenientes. Un individuo que se acerca a la religión de este modo reconoce que no sabe, y probablemente nunca sepa, la verdad última acerca de tales cuestiones. Aun así, las preguntas se consideran importantes y, a pesar de que son tentativas y sujetas a cambio, se buscan las respuestas. Puede existir o no una creencia clara en una realidad trascendente, pero existe un aspecto trascendental, religioso, en la vida del individuo. Le llamaremos a este enfoque abierto, de cuestionamiento, **religión como búsqueda** (Batson y cols., 1993, p.166).

Ejemplos de los 12 reactivos en la escala de Búsqueda son los siguientes:

- Mis experiencias vitales me han conducido a reconsiderar mis convicciones religiosas.
- Las dudas religiosas me parecen perturbadoras [se califica en sentido inverso].
- Espero que mis creencias religiosas cambien a medida que crezca y madure.
- Dentro de mi fe religiosa, las preguntas son más importantes que las respuestas.

Los reactivos se califican en el mismo continuo de nueve puntos desde (1) fuertemente en desacuerdo, hasta (9) fuertemente de acuerdo. Los resultados se presentan como una calificación promedio. La investigación con 424 estudiantes de pregrado interesados en la religión indica que la Búsqueda es, de hecho, una dimensión de la experiencia religiosa que es independiente tanto de la orientación Intrínseca como de la Extrínseca. Mientras que las puntuaciones de las escalas Intrínseca y Extrínseca se correlacionaron en .72, Búsqueda reveló relaciones insignificantes con ambas escalas (–.05 con Intrínseca y .16 con Extrínseca).

Pero ¿exactamente qué mide la escala de Búsqueda? La intención de sus autores fue que evaluara "el grado al que la religión de un individuo implica un diálogo abierto, receptivo, con preguntas existenciales surgidas de las contradicciones y tragedias de la vida" (p. 169). Los tres componentes de la orientación de búsqueda son: 1) disposición para enfrentar las preguntas existenciales sin reducir su complejidad, 2) autocrítica y percepción de las dudas religiosas como positivas y 3) apertura al cambio. Pero los críticos han presentado la acusación de que es posible que la escala no mida alguna especie de religiosidad y que en vez de ello evalúe posiblemente el agnosticismo, la actitud contraria a

la ortodoxia, las dudas religiosas o el conflicto religioso.

En respuesta a estas críticas, Batson y colaboradores (1993) señalan lo siguiente:

- Los estudiantes en el Seminario Teológico de Princeton tuvieron calificaciones significativamente más altas ($p < .001$) en la escala de Búsqueda (media de 6.7) que los estudiantes de pregrado en la misma institución (media de 5.2). Esta conclusión sustenta la perspectiva de que la escala es una medida válida de algún tipo de religiosidad.
- Los 32 miembros de un grupo carismático de estudio de la Biblia tuvieron calificaciones significativamente más elevadas ($p < .001$) en la escala de Búsqueda (media de 5.5), que los 26 miembros de un grupo tradicional de estudio de la Biblia (media de 4.6). El grupo carismático situaba el énfasis en la religión como búsqueda compartida; la mayoría rezaban con las manos en alto y algunos miembros hablaban en lenguas.

La Búsqueda es una dimensión independiente de expresión religiosa y se ha llevado a cabo una investigación sustancial acerca del significado y correlatos de esta orientación hacia la fe. Batson y colaboradores (1993) resumen la investigación con la escala de Búsqueda al señalar que parece medir una religión de menor fe, pero mayor esfuerzo.

El concepto de Búsqueda surgió en respuesta a las limitaciones del enfoque Intrínseco y Extrínseco para la medición de la orientación religiosa. Pero esta breve escala de 12 reactivos posee sus propias limitaciones, entre las cuales se pueden mencionar principalmente su brevedad y sencillez factorial. Se han propuesto varios otros instrumentos para medir aspectos de la experiencia religiosa. En las siguientes secciones se revisan unos cuantos enfoques destacados y representativos.

ESCALA DE BIENESTAR ESPIRITUAL

El concepto de bienestar espiritual puede rastrearse hasta un artículo de Moberg (1971) que propuso esta forma de bienestar como un componente esencial para el envejecimiento sano. El bienestar espiritual se conceptuó como un constructo bidimensional que consiste de una dimensión vertical y una horizontal. La primera tenía que ver con el bienestar en relación con Dios o un poder superior, mientras que la segunda implicaba un bienestar existencial que es un sentido de propósito en la vida sin ninguna referencia religiosa específica. Ellison (1983) y Paloutzian y Ellison (1982) emprendieron el reto de desarrollar una escala para medir estos componentes del bienestar.

Su instrumento se denominó la *Spiritual Well-Being Scale; SWB* (Escala de Bienestar Espiritual). La Escala SWB consiste de dos subescalas: Bienestar religioso (RWB), que evalúa la dimensión vertical del bienestar en relación con Dios; y Bienestar emocional (EWB), que mide la dimensión horizontal del bienestar en relación con el propósito y la satisfacción vitales. Cada subescala consiste de 10 reactivos que se califican desde 1 (fuertemente en desacuerdo) hasta 6 (fuertemente de acuerdo). Los reactivos de las dos subescalas se combinan en la Escala SWB, donde los reactivos nones evalúan el bienestar religioso y los pares examinan el bienestar existencial. Algunos de los reactivos se frasean en sentido negativo; éstos se califican de manera inversa, de modo que una calificación más alta siempre indica mayor bienestar. En el cuadro 12–8 se muestran reactivos similares a los encontrados en la Escala SWB.

La Escala SWB proporciona tres calificaciones: una calificación SWB total (máximo 120), una subcalificación para RWB (máximo 60) y una subcalificación para EWB (máximo 60). Los estudios iniciales de confiabilidad y validez se basaron en 206 estudiantes de tres universidades con orientación religiosa y una universidad secular. Los coeficientes de confiabilidad *test-retest* se colocaron en .93 para la SWB, .96 para la RWB y .86 para la EWB. Los análisis factoriales tendieron a sustentar la validez de constructo del instrumento al revelar que todos los reactivos religiosos tenían cargas en un factor religioso, mientras que los existenciales parecían tener carga en dos subfactores, uno

Cuadro 12–8. Reactivos similares a los encontrados en la Escala de bienestar espiritual

Encierre en un círculo la afirmación que mejor indique el grado de acuerdo o desacuerdo que usted tenga con ella

FA = Fuertemente de acuerdo
MD = Moderadamente de acuerdo
A = Acuerdo
D = Desacuerdo
MD = Moderadamente en desacuerdo
FD = Fuertemente en desacuerdo

No encuentro mucha recompensa en rezar de manera privada (calificación en dirección inversa)	FA	MD	A	D	MD	FD
Mi relación con Dios me ayuda a atravesar los momentos difíciles	FA	MD	A	D	MD	FD
La vida carece sustancialmente de significado (calificación en dirección inversa)	FA	MD	A	D	MD	FD
Me siento bien acerca de la ruta que sigue mi vida	FA	MD	A	D	MD	FD

vinculado con la dirección de la vida y el otro que indicaba satisfacción con la vida. La correlación entre las subescalas RWB y EWB fue modesta ($r = .32$), lo cual indica que detectan aspectos separados del bienestar espiritual.

En escritos posteriores, Ellison describió a la Escala SWB como una medida de la integración psicoespiritual de la personalidad y del bienestar resultante (Ellison y Smith, 1991). Según este punto de vista, el bienestar consiste de "la experiencia integral de una persona que funciona según los propósitos de Dios, en relación consonante con Él, con los demás y consigo misma" (p. 36). Éste es el concepto bíblico de *shalom*, el cual denota estar armoniosamente en paz dentro y fuera. Si esta conceptuación es correcta, la espiritualidad sana como la mide la Escala SWB podría mostrar relaciones positivas con medidas independientes de salud y bienestar subjetivo. Literalmente docenas de estudios han investigado esta hipótesis global, con conclusiones positivas en términos generales. En el cuadro 12–9 se resumen los estudios representativos.

El único defecto identificado de la Escala SWB es un límite superior aparentemente bajo,

Cuadro 12–9. Resumen de las conclusiones de la Escala de bienestar espiritual

Las calificaciones de la Escala de bienestar espiritual se correlacionan positivamente con:

Estar más cerca del peso corporal ideal (Hawkins y Larson, 1984).

Salud percibida en ancianos de áreas rurales (DeCrans, 1990).

Adaptación general a la hemodiálisis (Campbell, 1988).

Esperanza en pacientes con cáncer (Mickley, 1990).

Medidas de autoestima (Paloutzian y Ellison, 1982).

Las calificaciones de la Escala de bienestar espiritual se correlacionan negativamente con:

Presión sanguínea diastólica y sistólica (Hawkins, 1988).

Frecuencia y cantidad de dolor en pacientes con cáncer (Granstrom, 1987).

Aislamiento social y desesperación (Bonner, 1988).

Agresividad y evitación de conflictos (Bufford y Parker, 1985).

Calificaciones de depresión en el MMPI (Fehring, Brennan y Keller, 1987).

en especial con muestras religiosas. Ledbetter, Smith, Vosler-Hunter y Fischer (1991) advierten que la utilidad clínica de la escala está limitada a las calificaciones bajas (dado que las personas religiosas de alto funcionamiento tienden a "salirse del límite superior" de la escala). También ofrecen sugerencias para una revisión (p. ej., volver a frasear los reactivos en direcciones más extremas), con el propósito de elevar el nivel superior de la Escala SWB. Bufford, Paloutzian y Ellison (1991) han publicado normas para la prueba, pero advierten que en muchas muestras religiosas el individuo común recibe la calificación máxima. Esto indicaría que la escala es útil en investigación, pero no es adecuada para distinguir entre individuos con altos niveles de bienestar espiritual.

ESCALA DE MADUREZ DE LA FE

En 1987, seis de las principales denominaciones protestantes emprendieron un estudio nacional a cuatro años acerca de la fe personal, la alianza a una denominación y sus determinantes (Benson, Donahue y Erickson, 1993). Este proyecto que subvencionó en parte el *Lilly Endowment* abarcó lo que es sin duda la medición más compleja de la madurez espiritual que nunca se haya concebido. La *Faith Maturity Scale;* FMS (Escala de Madurez de la Fe) surgió como herramienta práctica para satisfacer tres propósitos de investigación:

1. Proporcionar una línea base sobre la vitalidad de la fe en las principales congregaciones protestantes.
2. Identificar las contribuciones de las variables demográficas, personales y de la congregación para el desarrollo de la fe.
3. Proporcionar un criterio variable para valorar el impacto de la educación religiosa en las principales denominaciones.

El desarrollo de la escala fue un proceso lento y cuidadoso que comenzó con una definición funcional:

> La madurez de la fe es el grado en que una persona abraza las prioridades, compromisos y perspectivas características de una fe activa y transformadora de la vida, como se le ha comprendido en las "principales" tradiciones protestantes (Benson, Donahue y Erickson, 1993, p. 3).

A través de cuestionarios abiertos con una muestra apropiada de 410 adultos de las principales denominaciones protestantes, los autores de la prueba identificaron a continuación ocho dimensiones centrales de la madurez de la fe. Tres páneles de asesores proporcionaron orientación continua durante esta etapa y en la fase siguiente de redacción de los reactivos. Estas interacciones garantizaron que la escala poseyera validez nominal y de contenido.

La resultante FMS es una prueba con 38 reactivos que representa los indicadores clave de la madurez de la fe en ocho áreas principales (cuadro 12–10). Los reactivos se responden en una escala de siete puntos desde 1 = nunca es cierto, hasta 7 = siempre es cierto. Con base en las áreas evaluadas, el lector notará que la creencia correcta es sólo un aspecto de la fe madura. En gran medida, la madurez de la fe se define por las consecuencias en cuanto a valores y conducta. Como señalan los autores, la Escala de madurez de la fe "se aparta de las maneras más tradicionales de definir y medir la religión personal". No obstante, representa los tipos de conductas y actitudes que se derivan de una fe dinámica y transformadora de la vida. Estas conductas y actitudes son consistentes con la teología que se encuentra en la mayor parte de las tradiciones religiosas, pero son especialmente pertinentes para el propósito de evaluar la madurez de la fe en el contexto protestante.

La FMS se califica como la media de los 38 reactivos, que produce un rango potencial de 1 a 7. La puntuación promedio de 3 040 adultos de cinco denominaciones protestantes fue de 4.63, lo cual indica que el instrumento evita el "efecto de límite superior" que se encuentra en otras escalas como la Escala de Bienestar Espiritual, que se analizó antes. La confiabilidad estimada de la escala es muy robusta entre edades, género, ocupación y denominación, con coeficientes alfa típicos de .88 (Benson *et al.*, 1993). No se presenta la confiabilidad *test-retest.*

Cuadro 12–10. Las ocho dimensiones centrales y reactivos de muestra de la Escala de madurez de la fe

A. Confía y cree (cinco reactivos)
Todos los días veo evidencia de la influencia de Dios en el mundo.

B. Experimenta los frutos de la fe (cinco reactivos)
Me siento abrumado por todas mis responsabilidades (calificación en dirección inversa).

C. Integra la fe y la vida (cinco reactivos)
Mi fe influye en la manera en que pienso y actúo todos los días.

D. Busca el crecimiento espiritual (cuatro reactivos)
Tomo tiempo para meditar o rezar.

E. Experimenta y nutre la fe en la comunidad (cuatro reactivos)
Hablo con otros acerca de mi fe.

F. Tiene valores que defienden la vida (seis reactivos)
Tiendo a ser crítico de otras personas (calificación en dirección inversa).

G. Defiende el cambio social (cuatro reactivos)
Creo que las iglesias de este país deberían participar en cuestiones políticas.

H. Actúa y sirve (cinco reactivos)
Ofrezco cantidades significativas de tiempo para ayudar a los demás.

Nota: Los reactivos de muestra son similares a los de la Escala de madurez de la fe.
Fuente: Basado en Benson, P., Donahue, M. y Erickson, J. (1993). The Faith Maturity Scale: Conceptualization, Measurement and Empirical Validation. En M. L. Lynn y D. O. Moberg (eds.), *Research in the social scientific study of religion* (vol. 5). Greenwich, CT: JAI Press.

La validez de la escala recibe apoyo de varias líneas de evidencia que comienzan con un cuidadoso enfoque para la selección de reactivos, a través del cual se incorporan la validez aparente y la de contenido. La validez de constructo se demostró de diversas maneras. Primero se predijo y se confirmó que los grupos supuestamente diferentes en niveles de madurez de la fe, obtendrían puntuaciones medias significativamente diferentes en la FMS. De hecho, los pastores tuvieron la puntuación más alta (5.3), seguidos de los coordinadores educativos de la iglesia (4.9), maestros (4.7), adultos (4.6) y jóvenes (4.1) —cada grupo en orden respectivo obtuvo calificaciones menores en términos significativos a los otros. En segundo lugar, las calificaciones de los pastores para la madurez de la fe de 123 miembros de la congregación en una escala de 1 a 10 se correlacionaron de manera bastante sustancial ($r = .61$) con las puntuaciones en la FMS de esas personas, lo cual indica una correspondencia entre las calificaciones independientes de un experto y el autoinforme. La escala también reveló utilidad predictiva. En términos específicos, las calificaciones de escala de la FMS se relacionaron en gran medida con una variedad de conductas prosociales como donar tiempo para ayudar a los pobres, hambrientos o enfermos; promover un papel más importante para las mujeres dentro de la iglesia; y respaldar el uso de la política exterior para desafiar la segregación racial.

Una advertencia es que la susceptibilidad de la escala a la propensión de las respuestas (p. ej., "fingirse bueno") simplemente se desconoce. Los autores de la prueba demandan una mayor investigación para examinar diversas propensiones de respuesta, como responder siempre sí y dar una apariencia socialmente deseable. También presentan una modestia refrescante al afirmar acerca de su prueba de 38 reactivos:

> Sería presuntuoso afirmar que estos 38 reactivos son la última palabra acerca de lo que define la fe madura. Esperamos que, a medida que pase el tiempo, la Escala de madurez de la fe también madure por medio de mayor información y modificación a través de la interacción con un extenso rango de investigadores y personas evaluadas que representen a un círculo cada vez más extenso de tradiciones religiosas (Benson y col., 1993, p. 24)

En general, la FMS representa una gran promesa como medida de la madurez de la fe. Una virtud especial de la prueba es que evita el efecto de límite superior tan común en otras medidas de las variables espirituales. Esta

característica garantiza que el instrumento resulte útil en el examen de los efectos de los programas educativos, así como de los cambios a lo largo del tiempo en individuos devotos.

ÍNDICE DE LA EXPERIENCIA ESPIRITUAL

Las medidas del funcionamiento espiritual y religioso se han desarrollado principalmente a partir de una interpretación cristiana de la fe, con el resultado de que las aplicaciones no pueden extenderse a personas cuyas creencias religiosas están enraizadas en otra ideología. Una notable excepción es el *Spiritual Experience Index* (SEI; Índice de experiencia espiritual), que se obtiene de una perspectiva del desarrollo de la fe. Esta prueba se basa en la teoría de que la fe es un fenómeno evolutivo que progresa de la religiosidad altamente egocéntrica de la infancia a la fe trascendente de la adultez media —cuando la nutren condiciones psicosociales favorables. Fowler (1981) fue el primero en dilucidar la teoría asociada con el desarrollo y después Genia (1990) la revisó. En el cuadro 12–11 se presenta un resumen de las cinco etapas de la fe según Genia.

Este nivel más elevado de madurez religiosa (que se encuentra en todas las religiones principales) se caracteriza por 10 criterios (Genia, 1993b):

1. Relación trascendente con algo mayor que uno mismo.
2. Consistencia del estilo de vida y conducta con los valores espirituales.
3. Compromiso sin certidumbre absoluta.
4. Apertura a puntos de vista espiritualmente diversos.
5. Ausencia de pensamiento mágico y de conceptos antropomórficos de Dios.
6. Inclusión de componentes tanto racionales como emocionales.
7. Interés social y preocupación humanitaria.
8. La fe madura mejora la vida y alienta el crecimiento.
9. Suministro de significado y propósito en la vida.
10. Falta de dependencia en prácticas particulares o estructura religiosa formal.

Cuadro 12–11. Etapas de la fe religiosa

Etapa 1. Fe egocéntrica

Característica de las personas con un desarrollo inmaduro de la personalidad, la fe egocéntrica es narcisista y se basa en la recompensa o castigo anticipados. La imagen divina es antropomórfica y la oración tiene propósitos de petición.

Etapa 2. Fe dogmática

El dogma religioso se utiliza de manera rígida y defensiva como medio de apoyo psicológico. Las escrituras se interpretan de modo literal y absoluto. La oración puede asumir la forma de convenio con Dios.

Etapa 3. Fe de transición

Característica de muchos adolescentes, el individuo reconoce una libertad para cuestionar y tener dudas. Puede ocurrir experimentación religiosa como "probar" diferentes ideologías o cambiar de denominación.

Etapa 4. Fe reconstruida internalizada

Existe un compromiso con una fe elegida por uno mismo que trasciende preocupaciones egocéntricas. La ideología elegida proporciona una sensación de propósito y significado en la vida. La doctrina religiosa es más compleja y la oración presenta agradecimiento, alabanza y devoción.

Etapa 5. Fe trascendente

El individuo mantiene una relación trascendente con algo más grande que él mismo. Existe un compromiso sin certidumbre absoluta y el estilo de vida de la persona es consistente con sus valores religiosos.

Fuente: Basado en Genia, V. (1990). Religious development: A synthesis and reformulation. *Journal of Religion and Health, 29*, 85-99.

El propósito de la escala SEI consiste en evaluar el grado de madurez espiritual (concebida como un constructo unidimensional) para las personas de diversas tradiciones religiosas y espirituales. Los reactivos de la escala se desarrollaron a partir de estos 10 criterios por medio de una construcción de escala racional (lo

cual significa que la autora de la prueba diseñó varios reactivos pertinentes para cada uno de los criterios). Una muestra de piloteo con 75 personas (40% católicos romanos, 28% protestantes, 23% judíos y 9% sin afiliación religiosa) respondió a una escala preliminar de 50 reactivos. Se eliminaron 12 reactivos debido a la baja correlación reactivo-total, lo cual dejó una escala de 38 reactivos.

Ejemplos de reactivos similares a los del SEI incluyen los siguientes:

- Mi fe me proporciona significado y propósito en la vida.
- En general un dilema moral sólo tiene una solución correcta. (Calificación en dirección inversa.)
- Siento una fuerte conexión espiritual con toda la humanidad.
- Mi fe me ayuda a enfrentar la tragedia y el sufrimiento.

El formato de los reactivos es una escala tipo Likert de seis puntos desde 1 = fuertemente en desacuerdo, hasta 6 = fuertemente de acuerdo. La calificación se presenta como la puntuación total natural con un rango de 38 a 228. Para la muestra inicial, las puntuaciones formaron una distribución normal que iba de 103 a 211, con una media de 166.

Se informa que la consistencia interna del SEI es de .87 (coeficiente alfa), pero no se evaluó la estabilidad *test-retest.* Cuando los 38 reactivos se sometieron a un Análisis de Factorización de Ejes de Principios, la escala se mantuvo como un constructo unidimensional —no surgieron factores que explicaran una parte significativa de la varianza. Por tanto, las conclusiones indican una buena confiabilidad interna para el SEI y sustentan su utilización como medida unidimensional.

La inspección de la validez consistió en análisis de correlación con una variedad de otras pruebas y medidas. Se encontraron fuertes correlaciones con diversas variables: dogmatismo ($r = -.52$), lo cual indica que las puntuaciones altas en la escala se asocian con bajo dogmatismo; puntuaciones de Búsqueda ($r = .44$), que sugieren que la escala incluye aspectos de complejidad y duda; religiosidad intrínseca ($r = .43$), lo cual indica que la fuerte experiencia espiritual promueve una fe religiosa madura; intolerancia de la ambigüedad ($r = -.40$), lo cual expresa que la fuerte experiencia espiritual permite la tolerancia de la ambigüedad; y frecuencia de la veneración ($r = .24$), lo cual es fascinante ya que ninguno de los reactivos de la SEI tiene que ver con este índice conductual de la fe.

En general, estos datos prestan sustentación a la confiabilidad y validez de la escala, pero se necesita de investigación adicional para examinar sus propiedades psicométricas. Es particularmente deseable un mayor análisis, en especial ya que el SEI ocupa un nicho especial en la evaluación de las variables espirituales —es una de las pocas escalas que puede aplicarse entre diversas tradiciones religiosas.

RESUMEN

1. Un valor es una creencia compartida y duradera acerca de los modos ideales de conducta o estados meta de la existencia. La evaluación de los valores es importante porque éstos persuaden a la acción, moldean actitudes y guían los esfuerzos para influir en los demás.
2. Uno de los primeros inventarios para evaluar los valores fue el Estudio de Valores, una prueba que consistía en 30 preguntas que contraponen un valor con otro y 15 preguntas que implican la clasificación de valores. Esta escala produjo un inventario de seis valores: Teóricos, Económicos, Estéticos, Sociales, Políticos y Religiosos.
3. El *Rokeach Value Survey* es otra herramienta para medir los valores. Esta prueba define y mide dos tipos de valores: instrumentales

(p. ej., Ambicioso, Imaginativo, Tolerante) y terminales (p.ej., una vida cómoda, armonía interna, una vida emocionante).

4. En la *Moral Judgement Scale* de Kohlberg, a la persona evaluada se le hace una serie de preguntas estructuradas relacionadas con varios dilemas morales. Las respuestas se categorizan según seis etapas y tres niveles de desarrollo: preconvencional, convencional y posconvencional.
5. La *Defining Issues Test (DIT)* de Rest es un derivado de la Escala de juicio moral que utiliza un formato de calificación completamente objetivo. La prueba produce varios índices cuantitativos, incluyendo la puntuación **P** (porcentaje de razonamiento de principios) y el índice N2 (basado en fórmulas complejas), ambos muestran buena confiabilidad y validez.
6. Un área relativamente descuidada en la evaluación es la valoración de las dimensiones espiritual y religiosa. Esto es desafortunado ya que la evaluación de estas variables puede tener importantes implicaciones para los clientes individuales.
7. Una de las primeras formas de evaluación religiosa fue la *Allport-Ross Religious Orientation scales*. Estas escalas popularizaron los conceptos de religiosidad intrínseca (las personas **viven** su religión para encontrar significado y dirección) contra religiosidad extrínseca (las personas **utilizan** su religión para buscar seguridad o estatus).
8. La escala *Quest* (Búsqueda) pretende medir una tercera orientación religiosa (más allá de la intrínseca y extrínseca) que se caracteriza por la complejidad, duda y vacilación como maneras de ser religioso. Esta sencilla escala de 12 reactivos parece medir una religión de menos fe, pero mayor esfuerzo.
9. La *Spiritual Well-Being Scale* es uno de los pilares en el campo de la evaluación religiosa. Consiste de dos subescalas: Bienestar religioso (una dimensión vertical de bienestar en relación con Dios) y Bienestar existencial (una dimensión horizontal de bienestar en relación con el propósito y la satisfacción en la vida). Un problema con esta escala es su bajo nivel superior.
10. El *Faith Maturity Index* es una ambiciosa escala diseñada por instrucciones de seis de las principales denominaciones protestantes. Los 38 reactivos se responden en una escala de siete puntos, lo cual proporciona un índice de una fe activa y transformadora de la vida, como la comprenden las principales iglesias protestantes.
11. El *Spiritual Experience Index* se diseñó para medir la madurez espiritual a partir de una perspectiva evolutiva de la fe, independiente de cualquier credo o religión particular. Los 38 reactivos se responden en una escala de seis puntos, lo cual produce un índice unitario de la madurez espiritual ampliamente definida.

TÉRMINOS Y CONCEPTOS CLAVE

Valor p. 555
Dilemas morales p. 560
Expresión religiosa intrínseca p. 565
Expresión religiosa extrínseca p. 565
Religión como Búsqueda p. 567

■

CAPÍTULO 13

Orígenes de las pruebas de personalidad

TEMA 13A
Teorías y medición de la personalidad

Personalidad: perspectiva general
Teorías psicoanalíticas de la personalidad
Teorías de tipos de personalidad
Teorías fenomenológicas de la personalidad
Teorías conductuales y del aprendizaje social
Conceptos sobre rasgos de personalidad
Resumen
Términos y conceptos clave

En las pruebas psigológicas, con frecuencia se establece una distinción fundamental entre las pruebas de capacidad y las de personalidad. Definidas en el sentido más amplio, las pruebas de capacidad incluyen la plétora de instrumentos para medir inteligencia, aprovechamiento, aptitud y funciones neuropsicológicas. En los 12 capítulos anteriores se ha explorado la naturaleza, construcción, aplicación, confiabilidad y validez de estos instrumentos, mientras que en los siguientes dos capítulos se cambia el énfasis hacia las pruebas de personalidad. Éstas miden uno o más de los siguientes aspectos: rasgos de personalidad, motivación dinámica, adaptación personal, sintomatología psiquiátrica, habilidades sociales y características actitudinales. Este capítulo investiga los orígenes de las pruebas de perso-

nalidad. En el tema 13A, Teorías y medición de la personalidad, se estudian las diferentes maneras en que los investigadores han conceptuado la personalidad para ilustrar la forma en que sus teorías han impactado al diseño de las pruebas y evaluaciones de la personalidad. En el tema 13B, Técnicas proyectivas, se analiza la multiplicidad de instrumentos basados en la hipótesis psicoanalítica de inicios del siglo XX de que las respuestas a estímulos ambiguos revelan los procesos mentales más internos, inconscientes, de la persona examinada. En el siguiente capítulo prosigue la cobertura de la evaluación de la personalidad con una revisión de las pruebas y procedimientos objetivos, incluyendo los inventarios de autoinforme y los enfoques de evaluación conductual.

PERSONALIDAD: PERSPECTIVA GENERAL

Aunque es difícil definir la personalidad, se puede distinguir dos características fundamentales de este constructo vago. Primero, cada persona es consistente en cierto grado; se tienen rasgos y patrones de acción coherentes que surgen de manera repetida. En segundo lugar, cada persona es distintiva hasta cierto punto; existen diferencias conductuales entre los individuos. Considérense las reacciones de tres estudiantes de posgrado cuando se les entregaron sus exámenes de mitad del curso. Aunque los tres estudiantes recibieron calificaciones casi idénticas (B), sus reacciones personales fueron bastante diversas. El primer estudiante se marchó malhumorado y después se le escuchó decir que sería adecuado poner una queja con el administrador del departamento. El segundo se mostró complacido, afirmando en voz alta que B era, después de todo, una calificación respetable. El tercero se sintió decepcionado, pero se mostró estoico. Se culpó a sí mismo por no estudiar con más ahínco.

¿Cómo se han de entender las diferentes reacciones de estas personas, cada una de las cuales respondía a un estímulo idéntico? Psicólogos y legos por igual invocan el concepto de **personalidad** para explicar la conducta y los sentimientos expresados por otros. Este concepto se utiliza para explicar las diferencias conductuales entre personas (por ejemplo, por qué uno se queja y otro se muestra estoico) y para comprender la consistencia dentro de cada individuo (por ejemplo, por qué el estudiante quejoso de quien se habló antes en general se mostraba amargado e insatisfecho).

Además de comprender la personalidad, los psicólogos también intentan medirla. Literalmente cientos de pruebas de personalidad están disponibles para este propósito; aquí se revisarán instrumentos históricamente destacados y también se analizarán algunos nuevos enfoques prometedores. No obstante, a fin de que el lector pueda comprender mejor la diversidad de instrumentos y enfoques, se comenzará con una pregunta fundamental: ¿Cómo puede conceptuarse mejor a la personalidad? Como descubrirá el lector, cuando se quiere medir la personalidad, primero debe preverse qué es lo que se busca medir. El lector apreciará mejor la multiplicidad de pruebas y procedimientos si también se describen de manera breve las teorías de personalidad que forman los fundamentos para estos instrumentos. Se cierra este tema con una pregunta general relacionada con todas las teorías y enfoques de pruebas: ¿Qué tan estable y predecible es la conducta?

Aunque se separan las pruebas de personalidad de las de capacidad, la distinción entre ambos tipos de instrumento está lejos de ser absoluta. La capacidad intelectual es, en parte, un rasgo caracterológico basado en atributos como perseverancia y autocontrol. Así, las pruebas de capacidad tocan inevitablemente importantes dimensiones de personalidad, aunque de manera indirecta e imperfecta. Con frecuencia también ocurre lo contrario; las pruebas de personalidad pueden estar saturadas de factores de capacidad. Por ejemplo, es probable que ciertas dimensiones de la personalidad, como la apertura a la experiencia, se correlacionen en sentido positivo con la inteligencia. Como el lector descubrirá en el siguiente capítulo, algu-

nos inventarios de personalidad con respuestas de cierto-falso incorporan un muy robusto factor de inteligencia (p. ej., Cattell, Eber y Tatsuoka, 1970).

TEORÍAS PSICOANALÍTICAS DE LA PERSONALIDAD

El psicoanálisis fue la creación original de Sigmund Freud (1856-1939). Aunque es cierto que muchos otros han revisado y adaptado sus teorías, los cambios han sido leves en comparación con los fundamentos sustanciales que pueden encontrarse de este genio singular de la era victoriana y de inicios del siglo XX. Freud fue sumamente prolífico en sus escritos y teorías. En este caso, se restringe este análisis a unos cuantos aspectos del psicoanálisis que han influido en las pruebas psicológicas. En particular, el Rorschach, el Test de apercepción temática y la mayor parte de las pruebas proyectivas revisadas en el siguiente tema requieren de un esquema psicoanalítico para su interpretación. Los lectores que deseen una revisión más amplia de las contribuciones de Freud pueden comenzar con Nuevas conferencias de introducción al psicoanálisis (Freud, 1933). Pueden encontrarse reseñas e interpretaciones de las teorías de Freud en Stafford-Clark (1971) y Fisher y Greenberg (1984).

ORÍGENES DE LA TEORÍA PSICOANALÍTICA

Freud comenzó su carrera profesional como neurólogo, pero pronto se especializó en el tratamiento de la histeria, un trastorno emocional caracterizado por conducta histriónica y síntomas físicos de origen psíquico como parálisis, ceguera y pérdida de sensación. Junto con su colega Joseph Breuer, Freud postuló que la raíz causal de la histeria eran las memorias sepultadas de experiencias traumáticas como el abuso sexual en la infancia. Si estos recuerdos pudieran recuperarse bajo hipnosis, podría ocurrir una liberación de la emoción llamada abreacción y los síntomas físicos desaparecerían, al menos por un breve tiempo (**Estudios sobre la histeria**, Breuer y Freud, 1893-1895).

A partir de estos primeros estudios, Freud desarrolló una teoría general del funcionamiento psicológico con el concepto de inconsciente como su fundamento. Consideraba que el inconsciente era el depósito de las pulsiones instintivas y un almacén de los pensamientos y deseos que serían inaceptables para nuestro yo consciente. Así, Freud discutía que nuestras motivaciones personales más importantes están en gran medida más allá de nuestra conciencia. El concepto de inconsciente se analizó con mayor detalle en su primer libro (**La interpretación de los sueños**, Freud, 1900). Freud creía que los sueños representan nuestros motivos inconscientes en una forma disfrazada. Incluso un sueño en apariencia inocuo podría tener en realidad un significado sexual o agresivo oculto, si se le interpreta de manera correcta.

El concepto de Freud acerca del inconsciente penetró hasta los meros fundamentos de las pruebas psicológicas a principios del siglo XX. A partir de allí surgió una familia completa de técnicas proyectivas, incluyendo manchas de tinta, enfoques de asociación de palabras, técnicas de Frases incompletas y de narración de historias (apercepción) (Frank, 1939; 1948). Cada uno de éstos se basaba en la suposición de que los motivos inconscientes pueden discernirse a partir de las respuestas de la persona examinada a estímulos ambiguos y desestructurados. De hecho, Rorschach (1921) relacionó su prueba de manchas de tinta con una radiografía de la mente inconsciente. Aunque exageró de manera patente el poder de las técnicas proyectivas, queda evidente, desde la perspectiva de Rorschach, que la concepción psicoanalítica del inconsciente tuvo una fuerte influencia sobre las prácticas de evaluación.

ESTRUCTURA DE LA MENTE

Los puntos de vista de Freud acerca de la estructura de la mente y la operación de los me-

canismos de defensa también influyeron en las pruebas y evaluación en psicología (**Nuevas conferencias de introducción al psicoanálisis**, Freud, 1933). Varias pruebas y enfoques de evaluación que se analizan en este capítulo se basan en el concepto psicoanalítico de los mecanismos de defensa, de modo que este tema merece un resumen.

Freud dividió a la mente en tres estructuras: ello, yo y superyó. El ello es la parte oscura e inaccesible de nuestra personalidad que Freud vinculó con "un caos, una caldera llena de excitaciones en ebullición". Debido a que el **ello** es totalmente inconsciente, se deben inferir sus características de manera indirecta a través del análisis de los sueños y de síntomas como la ansiedad. A partir de dicho análisis, Freud concluyó que el ello es el asiento de todas las necesidades instintivas como aquéllas de comida, agua, gratificación sexual y evitación del dolor. El ello tiene un solo propósito, obtener la satisfacción inmediata de estas necesidades de acuerdo con el **principio del placer**. Éste es el impulso hacia la satisfacción inmediata sin tomar en cuenta valores, bien o mal, o moral. El ello también es incapaz de lógica y no posee concepto de tiempo. Por tanto, los procesos mentales caóticos del ello permanecen inalterables con el paso del tiempo y las impresiones que se han guardado en éste "son virtualmente inmortales y se conservan durante décadas completas como si hubieran ocurrido apenas" (Freud, 1933).

Si nuestra personalidad consistiera sólo de un ello que se esfuerza por gratificar sus instintos sin tomar en cuenta la realidad, pronto seríamos aniquilados por las fuerzas externas. Por fortuna, poco después del nacimiento, parte del ello evoluciona para formar un **yo** o sí mismo consciente. El propósito del yo es mediar entre el ello y la realidad. El yo es parte y servidor del ello, pero "ha interpolado entre el deseo y la acción el factor aplazador del pensamiento". (Freud, 1933). Así, el yo es en gran medida consciente y obedece al **principio de realidad**; busca maneras realistas y seguras de descargar las tensiones instintivas que presionan constantemente desde el ello.

El yo también debe enfrentarse al **superyó**, el componente ético de la personalidad que comienza a surgir en los primeros cinco años de vida. Éste es casi un sinónimo de conciencia moral y comprende las normas sociales de lo que es correcto e incorrecto que nos trasmiten nuestros padres. El superyó es parcialmente consciente, pero gran parte de él es inconsciente; es decir, no siempre se está consciente de su existencia u operación. La función del superyó es restringir los intentos del ello y del yo de obtener gratificación. Su arma principal es la culpa, la cual utiliza para castigar las malas acciones del yo y del ello. Así, no es suficiente que el yo encuentre una manera segura y realista para gratificar las necesidades del ello; también debe elegir una salida moralmente aceptable o sufrirá el castigo de su supervisor, el superyó. Esto explica por qué uno puede sentirse culpable de conducta inmoral como el robo, aun cuando es imposible que se le atrape. Otra parte del superyó es el ideal del yo, que consiste de las metas y aspiraciones. El yo se mide contra el ideal del yo y lucha por satisfacer sus demandas de perfección. Si el yo no logra satisfacer los estándares del ideal del yo, el sentimiento de culpa puede ser el resultado. Por lo común se interpreta esta culpa como sentimiento de inferioridad (Freud, 1933).

PAPEL DE LOS MECANISMOS DE DEFENSA

Con toda seguridad el yo enfrenta una tarea difícil al actuar como mediador y servidor de tres tiranos: el ello, el superyó y la realidad externa. Al lector podría parecerle que la tarea sería esencialmente imposible y que, por tanto, el individuo se encontraría en un constante estado de angustia. Por fortuna, el yo tiene a su disposición una serie de herramientas para ayudarse a llevar a cabo su trabajo; a saber, las estrategias mentales que de manera colectiva se llaman **mecanismos de defensa.**

Éstos tienen muchas variedades, pero todos comparten tres características. Primero, su propósito exclusivo es ayudar al yo a reducir la

ansiedad creada por las demandas conflictivas del ello, superyó y realidad externa. De hecho, Freud consideraba que la ansiedad era una señal para que el yo invocara uno o más mecanismos de defensa en beneficio propio. Por tanto, los mecanismos de defensa y la ansiedad son conceptos complementarios en la teoría psicoanalítica que existen el uno como contrafuerza del otro. El segundo aspecto común de estos mecanismos es que operan a nivel inconsciente. Así, aunque están bajo el control del yo, uno no está consciente de su operación. Su tercera característica es que distorsionan la realidad interna o externa. Esta propiedad es la que los capacita para reducir la ansiedad. Al permitir que el yo vea de manera menos amenazante los desafíos del ello, del superyó o de la realidad externa, los mecanismos de defensa le ayudan a evitar niveles incapacitantes de ansiedad. Por supuesto, debido a que los mecanismos de defensa distorsionan la realidad, su aplicación rígida y excesiva puede crear más problemas de los que resuelve.

EVALUACIÓN DE LOS MECANISMOS DE DEFENSA Y DE LAS FUNCIONES DEL YO

Aunque Freud introdujo el concepto de mecanismos de defensa, sus seguidores fueron quienes dilucidaron en mayor detalle estas estrategias mentales inconscientes (Paulhus, Fridhandler & Hayes, 1997). Una de las primeras descripciones de los mecanismos de defensa la proporcionó la hija de Freud, Anna (**El yo y los mecanismos de defensa,** A. Freud, 1946). No obstante, la aplicación de estos conceptos a la medición y evaluación psicológicas es mucho más reciente. Por ejemplo, Loevinger (1976, 1979, 1984) ha producido una técnica de Frases incompletas para medir el desarrollo del yo que se basa, de manera indirecta, en el análisis de los mecanismos de defensa. En el siguiente tema se describe este interesante enfoque de la medición de la personalidad. Aquí se presentará el trabajo de Vaillant (1977, 1992) para ilustrar la medición de los mecanismos de defensa y la aplicación de esta información para la comprensión de la personalidad.

Vaillant (1971) desarrolló una jerarquía de mecanismos adaptativos del yo basada en la suposición de que algunos mecanismos de defensa son más sanos que otros. Desde esta perspectiva, los mecanismos de defensa pueden agruparse en cuatro tipos diferentes. Listados en orden de menor a mayor salud, los tipos son: psicóticos, inmaduros, neuróticos y maduros (cuadro 13–1). Los mecanismos psicóticos, como la negación absoluta de la realidad externa, son los menos sanos porque distorsionan la realidad en grado extremo. Al observador le parecen "locos". Los mecanismos de defensa inmaduros, como la proyección hacia otras personas de los propios sentimientos no reconocidos, son más sanos que los mecanismos psicóticos. No obstante, los observadores externos los detectan con facilidad y los consideran indeseables. Los mecanismos neuróticos de defensa por lo común alteran sentimientos privados, de modo que son menos amenazantes. Un ejemplo de esto es la intelectualización, un mecanismo de defensa en el que las cuestiones amenazantes se analizan en términos anodinos, vacíos de sentimiento. Por ejemplo, un médico cuya madre ha muerto recientemente podría hablar en gran detalle acerca de las características médicas del cáncer que ésta sufrió, con lo cual alivia su propio sentimiento de pérdida. Los mecanismos maduros de defensa al observador le parecen como virtudes convenientes. Un ejemplo lo constituyen ciertas formas de sentido del humor que no distorsionan la realidad pero que pueden aliviar la carga de asuntos "demasiado terribles como para tolerarlos" (Vaillant, 1977).

En el Estudio Grant se ilustra la aplicación de los mecanismos de defensa a la comprensión de la personalidad, un estudio de seguimiento con una duración de 45 años que condujeron Vaillant y otros colaboradores (Vaillant, 1977; Vaillant y Vaillant, 1990). Estos investigadores utilizaron entrevistas estructuradas para obtener evidencia de los mecanismos adaptativos inconscientes en una muestra de 95 varones. Los individuos provenían de una muestra original de 268 estudiantes de las clases de 1939 a 1944 en la Universidad de Harvard. En el seguimiento, Vaillant

Cuadro 13–1. Niveles de los mecanismos de defensa propuestos por Vaillant (1977)

I. Psicótico

Proyección delirante: francos delirios acerca de la realidad externa, en general de naturaleza persecutoria.

Negación: negación de la realidad externa; p. ej., falla en reconocer que se tiene una enfermedad terminal.

Distorsión: notoria reestructuración de la realidad externa para adecuarla a las necesidades internas; p. ej., delirios de cumplimiento del deseo.

II. Inmaduro

Proyección: atribuir los propios sentimientos no reconocidos a otras personas; p. ej. "¡Tú estás enojado, no yo!"

Fantasía esquizoide: uso de la fantasía y el retraimiento al interior con el propósito de resolver conflictos y obtener gratificación.

Hipocondria: transformación del reproche hacia los demás en autorreproche y después convertirlo en quejas de enfermedad física.

Conducta Agresiva-Pasiva: agresión hacia otros expresada indirecta e ineficazmente mediante pasividad o contra sí mismo.

Acting-out: expresión directa de un deseo o impulso inconscientes a fin de evitar tener conciencia del sentimiento que les acompaña.

III. "Neurótico"

Intelectualización: pensar sobre los deseos en términos formales, insensibles, pero no actuar sobre ellos.

Represión: fallas de memoria aparentemente inexplicables o fracaso en reconocer cierta información; p. ej., "olvidar" una cita con el dentista.

Desplazamiento: dirigir el sentimiento hacia algo o alguien que no es el objeto real; p. ej., patear al perro porque se siente enojo hacia el jefe.

Formación reactiva: cambiar de manera inconsciente un impulso en su contrario; p. ej., amabilidad exagerada hacia un compañero de trabajo a quien se odia.

Disociación: modificación temporal, pero drástica, del propio carácter para evitar angustia emocional; p. ej., una actitud de "me importa poco" por breve tiempo.

IV. Maduro

Altruismo: servicio vicario, pero constructivo y gratificante hacia los demás; p. ej., filantropía.

Sentido del humor: reconocimiento alegre de ideas y sentimientos sin incomodidad y sin causar efectos desagradables a los demás; no incluye el sarcasmo.

Supresión: decisión consciente o semiconsciente de posponer el prestar atención a un conflicto o impulso conscientes.

Anticipación: anticipación realista o planificación para cualquier incomodidad interna futura; p. ej., anticipación realista de una cirugía o una separación.

Sublimación: expresión indirecta de los deseos instintivos sin consecuencias adversas o pérdida del placer; p. ej., canalizar la agresión hacia los deportes.

Fuente: Basado en Vaillant, G. (1977). *Adaptation to life: How the best and the brightest came of age.* Boston, MA: Little, Brown.

entrevistó a cada uno de los participantes durante dos horas, utilizando un protocolo de entrevista semiestructurada (Vaillant, 1977, Apéndice B). Además, los participantes llenaron cuestionarios autobiográficos y proporcionaron otras fuentes de información. Vaillant y otros calificadores evaluaron los datos completos de cada persona según el grado en que cada mecanismo de defensa caracterizaba su adaptación vital. Los mecanismos de defensa se calificaron de 1 (ausente) a 5 (principal). A continuación se presenta un ejemplo de un mecanismo adaptativo inconsciente:

> Un hematólogo de California desarrolló el pasatiempo de cultivar células vivas en tubos de ensayo. En una entrevista reciente, describió con interés y animación especiales un cultivo inusualmente interesante que había obtenido de una biopsia de tejido proveniente de su madre. Sólo hacia el final de la entrevista reveló como por casualidad que su madre había muerto de una embolia apenas tres semanas antes. Su mención acerca de la muerte de su madre fue tan indiferente como su descripción acerca de cómo había logrado teñir de manera eficaz su cultivo de tejido aún vivo. De modo ingenioso e inconsciente, había utilizado su pasatiempo y sus habilidades especiales como médico para mitigar en forma temporal el dolor por su pérdida. Aunque su madre ya no estaba viva, al cambiar su atención todavía podía cuidarla. No había nada mórbido en la manera en que contaba la historia; y debido a que los mecanismos del yo son inconscientes, no tenía idea de su conducta defensiva. Muchos de los varones más sanos dentro del Estudio utilizaban tipos similares de cambios de atención o desplazamiento. A menos de que un observador entrenado la examine, dicha conducta pasa desapercibida la mayor parte de las veces (Vaillant, 1977).

Es muy probable que este individuo recibiera una calificación de 5 (principal) para el mecanismo de defensa neurótico de desplazamiento.

Si se toma en consideración el grado de juicio calificado que requería la tarea de evaluación, la confiabilidad intercalificadores para los mecanismos de defensa fue —con unas cuantas excepciones— respetable. Los mecanismos de defensa individuales poseían confiabilidades que iban de .53 (Fantasía) a .96 (Proyección); la mayor parte de las confiabilidades estaban en los rangos de .70 y .80. La confiabilidad de las calificaciones globales (que reflejaban la proporción de calificaciones de madurez a inmadurez) fue de .77.

La validez de las calificaciones para los mecanismos de defensa depende en gran medida de la demostración de que los cambios relacionados con el desarrollo y las diferencias grupales son consistentes con la teoría psicoanalítica en cuanto a estos constructos. Por ejemplo, se esperaría que los individuos del Estudio Grant utilizaran menos mecanismos de defensa inmaduros y más maduros a medida que alcanzaran la mediana edad y esto es precisamente lo que descubrió Vaillant. Además, se esperaría que las personas consideradas inadaptadas con base en otros criterios (p. ej., divorcio frecuente, desempeño deficiente) tuvieran calificaciones menos favorables en mecanismos de defensa en comparación con personas adaptadas, y esto también lo observó Vaillant. En suma, el análisis de los mecanismos de defensa es un enfoque prometedor para la evaluación de la personalidad. No obstante, este enfoque tiene dos inconvenientes: el examinador necesita entrenamiento especializado para reconocer los mecanismos de defensa y el proceso de obtención de la información pertinente de parte de los individuos examinados requiere mucho tiempo.

TEORÍAS DE TIPOS DE PERSONALIDAD

Las primeras teorías de personalidad intentaban clasificar a los individuos en categorías separadas o tipos. Por ejemplo, el médico griego Hipócrates (aproximadamente 460-377 a.C.) propuso una teoría sobre humores con cuatro tipos de personalidad (sanguíneo, colérico, melancólico y flemático), que era demasiado simplista como para ser útil. En el decenio de 1940, Sheldon y Stevens (1942) propusieron una teoría de tipos basada en la relación entre estructura corporal y temperamento. Su enfoque provocó una conmoción en la investigación

y después se desvaneció en el olvido. Sin embargo, las teorías tipológicas han continuado captando interés intermitente de parte de los investigadores de la personalidad. Se ilustrarán las teorías de tipo al revisar la investigación actual acerca de los tipos de personalidad con propensión a la cardiopatía.

PATRÓN CONDUCTUAL TIPO A, CON PROPENSIÓN A LA CARDIOPATÍA

Friedman y Rosenman (1974) investigaron las variables psicológicas que colocaban a los individuos en mayor riesgo de cardiopatía coronaria. Fueron los primeros en identificar un **Patrón conductual Tipo A con propensión a la cardiopatía**, que describieron como un "complejo acción-emoción que puede observarse en cualquier persona que participa de manera agresiva en una lucha crónica, incesante, para lograr cada vez más en menor tiempo y que, si se necesita, lo hace en contra de los esfuerzos opuestos de parte de otras cosas o personas" (Friedman y Rosenman, 1974). En el otro extremo se encuentra el patrón conductual Tipo B, que se describe como un tipo de vida despreocupado, no competitivo y relajado. Por supuesto, las personas varían a lo largo de un continuo desde el Tipo A "puro" hasta el Tipo B "puro".

Friedman y Ulmer (1984) han listado los componentes específicos de un patrón conductual Tipo A completo:

- **Inseguridad acerca del estatus:** muchas personas Tipo A parecen estar afligidas por una oculta falta de autoestima. No importa qué tan exitosas sean, con frecuencia se comparan de manera desfavorable con otras personas con logros altos.
- **Hiperagresividad:** parte de este patrón lo constituye un deseo de dominar a los demás y dañar su autoestima. Con frecuencia, las personas Tipo A se muestran indiferentes a los sentimientos o derechos de sus competidores.
- **Hostilidad libre flotante:** la persona Tipo A tiene muchas cosas por las cuales molestarse y el enojo está fuera de proporción con la situación.
- **Sentido de urgencia de tiempo (enfermedad de la prisa):** esto incluye dos estratagemas básicas: acelerar las actividades diarias (¡una persona Tipo A utilizaba una rasuradora eléctrica en cada mano!) y hacer dos cosas a la vez, como conversar por teléfono mientras se revisa la correspondencia.

La conducta Tipo A puede diagnosticarse a partir de una breve entrevista que consiste de preguntas acerca de hábitos para trabajar, hablar, comer, leer y pensar (Friedman y Ulmer, 1984). También pueden detectarse los casos más evidentes de conducta Tipo A a través de pruebas de lápiz y papel (Jackson y Gray, 1987; Jenkins, Zyzanski y Rosenman, 1971, 1979) que se analizarán en el siguiente capítulo. No obstante, el enfoque de cuestionario es limitado, debido a que no puede revelar los índices faciales, vocales y psicomotores de hostilidad y urgencia de tiempo que por lo general son evidentes en la entrevista (Friedman y Ulmer, 1984).

Los primeros estudios indicaron que las personas que exhibían el Patrón Conductual Tipo A estaban en mucho mayor riesgo de sufrir cardiopatía coronaria e infarto al miocardio. En un estudio de nueve años con más de 3 000 varones sanos, las personas con patrón conductual Tipo A estaban en una probabilidad 2.5 veces mayor de sufrir un infarto que aquéllos con un patrón conductual Tipo B (Friedman y Ulmer, 1984). De hecho, ninguno de los individuos con Tipo B "puro" —aquellos miembros del estudio que eran muy relajados, despreocupados y no competitivos— sufrió un infarto. En el famoso estudio longitudinal Framingham, los varones Tipo A entre 55 y 64 años de edad mostraron una probabilidad dos veces mayor de desarrollar cardiopatía, al momento del seguimiento a los 10 años, que aquéllos del Tipo B (Haynes, Feinleib y Eaker, 1983). En este estudio, la vinculación entre la conducta Tipo A y la cardiopatía fue especialmente fuerte entre trabajadores no obreros.

En estudios más recientes, los investigadores han encontrado sólo una relación leve

—o ninguna relación en absoluto— entre la conducta Tipo A y la cardiopatía coronaria (p. ej., Eaker y Castelli, 1988; Mathews y Haynes, 1986). Otros investigadores han encontrado que la cardiopatía no se liga tanto con el patrón conductual Tipo A completo como con sus componentes específicos, como la propensión al enojo (Dembroski, MacDougall, Williams y Haney, 1985). Seguramente, existe la necesidad de clasificar los factores específicos de riesgo en esta área de investigación. En Brannon y Feist (1992) y Wiebe y Smith (1997) pueden encontrarse buenas revisiones acerca de la compleja y confusa investigación sobre la conducta Tipo A.

La investigación sobre la conducta Tipo A ha despertado un interés renovado y más complejo hacia los conceptos tipológicos de la personalidad. En vez de ver los tipos como casillas separadas, los psicólogos han llegado a considerarlos como ejemplos idealizados que ocupan los puntos terminales de dimensiones continuas. Así, los individuos pueden diferir con respecto a qué tanto poseen de un tipo de personalidad idealizado. Esto es similar al concepto de rasgos de personalidad que se analiza después. Quizá la diferencia principal es que, en la actualidad, los teóricos de los tipos tienden a creer que la mayoría de los individuos se encuentran cerca de los tipos idealizados al final de cada dimensión, mientras que los teóricos de los rasgos afirman que es más probable que las personas se coloquen en todos los puntos a lo largo de cada continuo de personalidad. Por tanto, en la práctica, la distinción moderna entre tipos y rasgos es relativa, no absoluta.

TEORÍAS FENOMENOLÓGICAS DE LA PERSONALIDAD

Las teorías fenomenológicas de la personalidad enfatizan la importancia de la experiencia inmediata, personal y subjetiva como determinante de la conducta. Algunas de las posturas teóricas resumidas bajo este título han recibido también otros nombres, como teorías humanistas, existenciales, de constructo, del sí mismo y de realización (Maddi, 1989). Sin embargo, estos enfoques comparten la atención que dirigen a la experiencia subjetiva de la persona, la perspectiva personal acerca del mundo y el autoconcepto como fuentes principales de la conducta.

ORÍGENES DEL ENFOQUE FENOMENOLÓGICO

La orientación que se revisa de manera breve en esta sección tiene numerosas fuentes que llegan hasta la filosofía y literatura europeas de principios del siglo XX. No obstante, dos personas, uno filósofo y el otro escritor, se destacan como los contribuyentes originales al punto de vista fenomenológico moderno. El filósofo alemán, Edmund Husserl (1859-1938), inventó la compleja filosofía de la fenomenología que se ocupaba de la descripción de los fenómenos mentales puros. Su enfoque era principalmente introspectivo y casi inescrutable. Más asequible era el escritor danés Soren Kierkegaard (1913-1855), muy conocido por sus contribuciones al existencialismo. Éste es el movimiento literario y filosófico que se ocupa del significado de la vida y de la libertad del individuo para elegir sus propias metas. La fenomenología de Husserl y el existencialismo de Kierkegaard influyeron a docenas de importantes filósofos y psicólogos. Los vestigios de estos primeros puntos de vista son evidentes en casi cualquier teoría fenomenológica contemporánea (Maddi, 1989).

CARL ROGERS, LA TEORÍA DEL SÍ MISMO Y LA TÉCNICA Q

El teórico fenomenológico con mayor influencia fue Carl Rogers (1902-1987). Sus contribuciones a la teoría de la personalidad, conocida como teoría del sí mismo, son extensas y en general muy bien aceptadas por los estudiantes de psicología (Rogers, 1951, 1961, 1980). Pero también es cierto, aunque sea poco reconocido, que Rogers ayudó a conformar una peque-

ña parte de las pruebas psicológicas al popularizar la técnica Q.

La **técnica Q** es un procedimiento para estudiar los cambios en autoconcepto, un elemento clave en la teoría del sí mismo de Rogers. Fue Stephenson (1953) quien desarrolló la técnica, pero una serie de estudios realizados por Rogers y sus colaboradores sirvió para popularizar este enfoque de medición (Rogers y Dymond, 1954). La técnica Q, también conocida como Clasificación Q, es un procedimiento generalizado que es especialmente útil para estudiar los cambios en el autoconcepto.* La clasificación Q consiste de un gran número de tarjetas, cada una de las cuales contiene una afirmación impresa como las siguientes:

Soy equilibrado.
Doy una apariencia falsa.
Me hago fuertes demandas.
Soy una persona sumisa.
Soy agradable.

Se pide a la persona examinada que clasifique aproximadamente 100 de las afirmaciones en nueve pilas, colocando un número determinado de tarjetas en cada una de ellas, con lo cual se fuerza una distribución casi normal. Las instrucciones especifican que la persona debe colocar en un extremo las tarjetas que más la describen, en el extremo opuesto deben ir aquellas que menos la describen y aquéllas hacia las cuales se muestra indiferente o indecisa deben colocarse a la mitad de la distribución. Ésta podría parecerse más o menos a la siguiente:

	Menos parecido a mí								**Más parecido a mí**
Núm. de pila	1	2	3	4	5	6	7	8	9
Núm. de tarjetas	1	4	11	21	26	21	11	4	1

La naturaleza de los reactivos se determina por las necesidades del investigador o del profesional. Rogers utilizaba un conjunto de reactivos diseñados por Butler y Haigh (Rogers y Dymond, 1954, cap. 4) para llegar al autoconcepto. Estas afirmaciones se tomaron al azar de protocolos terapéuticos disponibles; sus reactivos en la clasificación Q representaban afirmaciones reales de clientes, parafraseadas con claridad. Pero una virtud especial de la técnica Q es que otros investigadores o profesionales están en libertad de diseñar sus propios reactivos. Por ejemplo, Marks y Seeman (1963) utilizaron una perspectiva psicodinámica para diseñar reactivos para las descripciones del terapeuta acerca de grupos de pacientes. Ejemplos de éstos incluyen:

Utiliza *acting out* como mecanismo de defensa.
Tiende a ser indiferente tanto de palabra como de obra.
Genotipo tiene características paranoides.
Parece equilibrado, seguro de sí mismo, socialmente relajado.
Exhibe depresión (tristeza manifiesta).

En general, la calificación que se asigna a una clasificación Q es cuestión de comparar o correlacionar la distribución de reactivos contra una norma establecida. Por ejemplo, podría pedirse a personas bien adaptadas que clasificaran los reactivos de modo que se obtuviera un número promedio de colocación por pila (de 1 al 9) para cada reactivo. A un individuo examinado se le consideraría más o menos adaptado según el parecido entre sus clasificaciones y la clasificación promedio de personas adaptadas. El lector puede consultar a Block (1961) para los detalles.

Otra manera de utilizar la clasificación Q consiste en comparar la autoclasificación de la persona evaluada con su clasificación ideal. Rogers utilizaba la discrepancia entre ambas clasificaciones como índice de adaptación. Pedía a sus personas que clasificaran los reactivos en dos ocasiones, de acuerdo con las siguientes instrucciones:

1. **Autoclasificación.** Clasifique estas tarjetas para describirse a usted mismo como se ve

* La técnica Q tiene también aplicaciones adicionales. Marks y Seeman (1963) emplearon clasificaciones Q de terapeutas para describir a pacientes con perfiles específicos en el MMPI. Bem y Funder (1978) recomiendan una clasificación Q para obtener un perfil de características asociadas con el desempeño exitoso en una tarea específica. Puede predecirse el éxito en la tarea seleccionada de las personas cuya autodescripciones encajan en el perfil obtenido.

hoy día, desde aquellas que representan menos lo que es usted hasta aquellas que se parecen más a usted.

2. **Clasificación ideal.** Ahora clasifique estas tarjetas para describir a su persona ideal —la persona que usted, en su interior, más quisiera ser (Rogers y Dymond, 1954).

Al utilizar los números de pila de los reactivos, Rogers correlacionaba por separado las dos clasificaciones de cada individuo. Considérese lo que significan estos datos: si la autoclasificación y la clasificación ideal son sumamente similares, la correlación de los datos de la clasificación Q se acercarán a 1.0; si ambas clasificaciones son opuestas, la correlación se acercará a –1.0. Por supuesto, la mayor parte de las clasificaciones se colocarán en algún sitio intermedio, pero por lo común hacia el lado positivo. Butler y Haigh encontraron que los clientes en psicoterapia aumentaban su congruencia entre su propia imagen y su ideal (Rogers y Dymond, 1954, cap. 4). Aun así, los participantes control adaptados poseían mayor congruencia (cuadro 13–2).

TEORÍAS CONDUCTUALES Y DEL APRENDIZAJE SOCIAL

Las teorías conductuales y del aprendizaje social tienen sus orígenes en estudios de laboratorio sobre el aprendizaje operante y el condicionamiento clásico. Una suposición fundamental de todos los teóricos conductuales es que muchas de las conductas que conforman la personalidad son aprendidas. Entonces, para comprender la personalidad, se debe conocer la historia de aprendizaje del individuo. Los teóricos conductuales también consideran que el ambiente tiene una importancia suprema en el moldeamiento y mantenimiento de la conducta. Por tanto, la investigación conductual busca identificar los componentes dentro del ambiente actual que controlan la conducta de una persona. El enfoque conductual de la personalidad ha producido una variedad de métodos de evaluación directa que se analizan en el siguiente capítulo.

El desacuerdo principal de los teóricos conductuales se refiere al papel que representan las cogniciones como determinantes de la conducta. Éstas son procesos mentales inferidos como la solución de problemas, el juicio o el razonamiento. Los conductistas radicales consideran que es inútil recurrir a explicaciones mentalistas de algún tipo: "Cuando aquello que hace una persona se atribuye a lo que ocurre en su interior, la investigación llega a su fin" (Skinner, 1974). En contraste, los teóricos del aprendizaje social hacen una referencia cauta a las cogniciones en la explicación de qué es, en términos específicos, lo que una persona aprende. Un teórico del aprendizaje social podría discutir que aprendemos expectativas o reglas acerca del ambiente, no sólo conexiones entre estímulo y respuesta.

En la actualidad, la teoría del aprendizaje social puede considerarse como una variante cognitiva del conductismo estricto que dominaba a la psicología estadounidense a principios del siglo XX. Los teóricos del aprendizaje social aceptan la premisa skinneriana de que el reforzamiento externo es un determinante importante de la conducta. Pero a la vez sostienen que las cogniciones también tienen una influen-

Cuadro 13–2. Correlaciones promedio del sí mismo ideal para los grupos de clientes y control

	Preorientación	Posorientación	Seguimiento
Grupo de clientes ($N = 25$)	–.01	.36	.32
Grupo control ($N = 16$)	.58		.59

Fuente: Basado en Rogers, C. R. y Dymond, R. F. (eds.) (1954). *Psychotherapy and personality change: Co-ordinated research studies in the client-centered approach.* Chicago: The University of Chicago Press.

cia crítica sobre nuestros actos. Por ejemplo, Rotter (1972) ha popularizado el punto de vista de que nuestras expectativas acerca de los resultados futuros son los principales determinantes de la conducta. Por ejemplo, la probabilidad de que una persona se comporte de manera agresiva depende de sus expectativas acerca de los resultados probables de su propia agresividad. Si la persona valora el resultado esperado, la conducta se vuelve más probable. Por supuesto, las expectativas están en función de la historia de reforzamiento de la persona, de modo que la perspectiva del aprendizaje social de Rotter es similar al punto de vista conductual. Pero la implicación de la teoría del aprendizaje social es que la conducta es el resultado de una creencia, en particular aquélla de que la conducta tendrá como efecto un resultado deseado. Así, se supone que las cogniciones afectan a las acciones.

Con base en sus perspectivas del aprendizaje social, Rotter (1966) desarrolló la Escala interna-externa (I-E), una interesante medida del *locus* de control interno contra el externo. El constructo del ***locus* de control** se refiere a las percepciones que tienen los individuos acerca de la fuente de las cosas que les suceden. En particular, la Escala I-E busca evaluar las expectativas generalizadas de la persona en cuanto al control de reforzamiento interno contra el externo. El propósito de la Escala I-E consiste en determinar el grado en que la persona examinada cree que el reforzamiento es contingente a su conducta (*locus* de control interno) más que dependa del mundo exterior (*locus* de control externo). El instrumento es un inventario de autoinforme de elección forzada. Para cada reactivo, el individuo evaluado elige la única afirmación (de un par) con la que más coincide. Éstos se parecen a los siguientes:

> En general, la mayoría de las personas reciben el respeto que se merecen.
>
> O
>
> En realidad, es frecuente que la valía de una persona pase desapercibida.

Para el reactivo anterior, la primera alternativa indica un *locus* de control interno, mientras que la segunda implica un *locus* de control externo. El equilibrio entre respuestas internas y externas determina la calificación total en la escala. La escala I-E es un instrumento confiable y válido que ha estimulado una amplia cantidad de investigaciones acerca de la naturaleza y significado del *locus* de control y variables relacionadas. La investigación indica que el locus de control tiene una fuerte relación con el éxito laboral, la salud física, logros académicos y otras numerosas variables. Como podría sospechar el lector, un *locus* de control interno en general predice un resultado más positivo que un *locus* de control externo. El lector interesado puede consultar a Lefcourt (1982) y Wall, Hinrichsen y Pollack (1989) para mayores detalles.

Albert Bandura también ha realizado importantes contribuciones a la teoría del aprendizaje social. En sus primeros estudios, Bandura examinó el papel del aprendizaje observacional y del reforzamiento vicario en el desarrollo de la conducta (Bandura, 1965, 1971; Bandura y Walters, 1963). De manera más reciente, ha propuesto que la autoeficacia percibida es un mecanismo central en las acciones humanas (Bandura, 1982; Bandura, Taylor, Ewart, Miller y DeBusk, 1985). La **autoeficacia** es un juicio personal acerca de "qué tan bien puede uno llevar a cabo los cursos de acción requeridos para manejar situaciones eventuales" (Bandura, 1982). El concepto de autoeficacia es útil en la explicación de por qué el conocimiento correcto no necesariamente predice una acción eficaz. Por ejemplo, dos niños pueden estar convencidos por igual de que una serpiente de jardín en la tina del baño no representa un peligro, pero uno la recoge mientras que el otro sale corriendo. Estas diferencias en conducta ilustran el papel del pensamiento autorreferencial como mediador entre conocimiento y acción. El niño que sale corriendo no creía que pudiera manejar la situación de manera eficaz. Tenía poca autoeficacia percibida en cuanto a manejo de serpientes. Bandura discutiría que el determinante principal de la conducta del niño es su propio juicio acerca de sus capacidades personales. Por tanto, se supone que las cogniciones son un determinante primordial de la conducta.

Bandura ha desarrollado un interesante instrumento para la evaluación de las expectativas de autoeficacia (Bandura, Taylor, Ewart, Miller y BeBusk, 1985). Para una variedad de situaciones que podrían excitar ansiedad, molestia o enojo, el individuo examinado marca si "puede hacer" la tarea y también califica el grado de confianza utilizando un número de 10 a 100. El formato de la lista de verificación es como sigue:

10 20 30 40 50 60 70 80 90 100

Bastante inseguro — Moderadamente seguro — Seguro

	Puedo hacerlo	Confianza
Ir a una fiesta en la que no haya nadie conocido.	_____	_____
Quejarse de la mala comida en un restaurante.	_____	_____

En esencia, el instrumento de Bandura es una herramienta referida a criterio que se utiliza en psicoterapia e investigación.

CONCEPTOS SOBRE RASGOS DE PERSONALIDAD

Un **rasgo** es cualquier "manera relativamente duradera en la que un individuo difiere de otro" (Guilford, 1959). Los psicólogos desarrollaron el concepto de rasgo a partir de las formas en que las personas describen a otras en la vida cotidiana. A medida que evolucionó el lenguaje, las personas encontraron palabras para representar las consistencias y diferencias que enfrentaban en sus interacciones diarias con los demás. Así, cuando se dice que una persona es sociable y otra es tímida, se utilizan nombres de rasgos para describir las consistencias dentro de los individuos y también las diferencias entre ellos (Goldberg, 1981a; Fiske, 1986).

Los conceptos sobre rasgos de personalidad han sido sumamente populares a lo largo de la historia de las pruebas psicológicas, de modo que en nuestro caso la cobertura será selectiva por necesidad. Se revisarán tres posturas importantes e influyentes entre las docenas de teorías de rasgos que se han propuesto. Estos enfoques difieren principalmente en términos de si los rasgos se dividen en variantes finamente discriminables o se agrupan en un pequeño número de dimensiones amplias:

1. El punto de vista analítico-factorial de Cattell identifica de 16 a 20 dimensiones bipolares de rasgo.
2. El enfoque dimensional de rasgos de Eysenck une docenas de rasgos en dos dimensiones prevalecientes.
3. Goldberg y otros han buscado una síntesis moderna de todos los enfoques de rasgo al proponer un modelo de los cinco factores de la personalidad.

Para los lectores que deseen un análisis más detallado sobre este tema, Pervin (1993) y Wiggins (1997) proporcionan una excelente reseña de los enfoques de rasgo dentro de la teoría de personalidad.

TEORÍA ANALÍTICO-FACTORIAL DE RASGOS DE CATTELL

Cattell (1950, 1973) depuró los métodos de análisis factorial existentes para ayudar a revelar los rasgos básicos de la personalidad. Denominó **rasgos superficiales** a los aspectos más obvios de la personalidad. Por lo común, éstos surgirían en las primeras etapas del análisis factorial cuando se correlacionaran entre sí los reactivos individuales de una prueba. Por ejemplo, reactivos de cierto-falso como "Disfruto una buena pelea de campeonato", "Realmente me molesta quedar atrapado detrás de un conductor lento" y "Es importante hacerle saber a la gente que uno tiene el mando", podrían recibir respuestas similares de parte de los individuos, con lo cual se revelaría un rasgo superficial de agresividad.

No obstante los rasgos superficiales en sí tendían a agruparse, como lo revelaron las apli-

caciones más complejas del análisis factorial que realizó Cattell. Para él, ésta era evidencia de la existencia de **rasgos fuente**, los orígenes estables y constantes de la conducta. Por tanto, los rasgos fuente son menos visibles que los superficiales, pero son más importantes en la explicación de la conducta.

Cattell (1950) no tuvo rival en la utilización del análisis factorial para descubrir la manera en que se organizaban los rasgos y cómo se relacionaban entre sí. Un enfoque consistía en hacer que las personas calificaran a otros individuos a quienes conocían, al marcar diversos adjetivos como **agresivo**, **considerado** y **dominante** en una lista de 171 opciones. Cuando posteriormente se analizaron factorialmente los resultados de 208 personas, se identificaron de modo tentativo cerca de 20 factores o rasgos subyacentes de personalidad. Otro enfoque consistió en hacer que miles de personas respondieran preguntas acerca de sí mismas y después realizar un análisis factorial de sus respuestas. A través de este segundo enfoque, se confirmaron de manera independiente 16 de los 20 rasgos originales de personalidad (Cattell, 1973). Estos 16 rasgos fuente se han incorporado al **Cuestionario de 16 factores de personalidad**, una prueba de lápiz y papel para los rasgos de personalidad que se analiza en el siguiente capítulo.

TEORÍA DIMENSIONAL DE RASGOS DE EYSENCK

Eysenck utilizó el análisis factorial para producir un acercamiento económico entre los enfoques de rasgo y dimensión de la personalidad (Eysenck y Eysenck, 1975, 1985). Según su sistema, la personalidad consiste de dos dimensiones básicas, introvertida-extravertida y emocionalmente estable-emocionalmente inestable. Se supone que estas dos dimensiones tienen una base biológica y genética. Además, las dimensiones incluyen numerosos rasgos específicos (figura 13–1). Las posiciones de los 32 rasgos corresponden a la dirección y cantidad de las dos dimensiones básicas. Por ejemplo, una persona moderadamente extravertida que sea inestable a nivel moderado podría caracterizarse por los siguientes rasgos: agresiva, excitable, cambiante. Una persona muy introvertida que se encuentre a la mitad en la dimensión estable-inestable podría considerarse como poco sociable, tranquila, pasiva y cuidadosa. La teoría dimensional de rasgos de personalidad de Eysenck se incorpora en su inventario de personalidad, el Cuestionario de personalidad de Eysenck, que se revisará en el siguiente capítulo.

MODELO DE LOS CINCO FACTORES DE LA PERSONALIDAD

El modelo de los cinco factores de la personalidad tiene sus orígenes en una revisión realizada por Goldberg (1981b). En su análisis de la investigación analítico-factorial de los rasgos, Goldberg identificó varias consistencias a las que denominó las "Cinco grandes" dimensiones. Aunque los investigadores han utilizado términos un tanto diferentes para estos factores, los nombres más comunes son:

Neuroticismo.
Extraversión.
Apertura a la experiencia.
Conformidad.
Escrupulosidad.

Un reacomodo de los factores produce las sencillas siglas: NACEE. El modelo de los cinco factores se ha vuelto rápidamente el modelo consensual de la personalidad. La sustentación para el enfoque de los cinco factores proviene de varias fuentes que incluyen el análisis factorial de los términos para rasgos dentro del lenguaje y el análisis de la personalidad a partir de una perspectiva evolutiva. A continuación se analizan estos puntos de vista.

El uso de los términos de rasgo en el análisis de la personalidad se basa en la **hipótesis léxica fundamental**. El punto esencial de esta hipótesis es que los términos para rasgos han sobrevivido en el lenguaje debido a que trasmiten información importante acerca de nuestro trato con otras personas:

> La variedad de diferencias individuales es casi ilimitada; sin embargo, la mayor parte de estas



Figura 13–1. Clasificación bidimensional de los rasgos de personalidad de Eysenck.
Fuente: Reproducido con autorización de Eysenck, H. J. y Eysenck, M.W. (1985). *Personality and individual differences: A natural science approach.* New York: Plenum.

> diferencias son insignificantes en las interacciones cotidianas de las personas con otros individuos y casi nunca se les reconoce. Sir Francis Galton puede haber estado entre los primeros científicos en reconocer de manera explícita la hipótesis léxica fundamental —es decir, que la mayor parte de las diferencias individuales importantes en las transacciones humanas se codificarán con términos sencillos en algunos o en todos los idiomas del mundo (Goldberg, 1990).

Cuando en el caso del idioma inglés los términos de rasgo se reducen a un conjunto razonablemente concreto y no superpuesto de adjetivos, por lo común surgen unos cuantos cientos de características (Allport, 1937). Durante decenios, los investigadores le han pedido a los individuos que se califiquen a sí mismos o a otras personas en éstos u otros rasgos similares. Cuando estas calificaciones se someten a un análisis factorial, por lo general aparecen de una manera u otra las cinco grandes dimensiones que se señalan antes. En resumen, un conjunto cada vez mayor de investigaciones indica que el modelo de los cinco factores capta una representación válida y útil de la estructura de los rasgos humanos.

El enfoque de los cinco factores también posee credibilidad evolutiva. En específico, los cinco factores de personalidad que se mencionaron antes captan diferencias individuales que se relacionan con funciones evolutivas básicas como la supervivencia y el éxito reproductivo (Buss, 1997; Pervin, 1993). Goldberg (1981b) ha teorizado que las personas hacen de manera implícita las siguientes preguntas en sus interacciones con otros:

1. ¿X es activo y dominante o pasivo y sumiso? (¿Puedo intimidar a X o X tratará de intimidarme?)

2. ¿X es afable (cálido y agradable) o desagradable (frío y distante)?
3. ¿Puedo contar con X? (¿X es responsable y escrupuloso o poco confiable y negligente?)
4. ¿X está loco (imprevisible) o cuerdo (estable)?
5. ¿X es listo o tonto? (¿Qué tan fácil me será enseñar a X?)

De manera directa o indirecta, cada una de estas valoraciones se relaciona con la supervivencia o el éxito reproductivo. Por ejemplo, el punto 3 (escrupulosidad) implica un rasgo que podría asegurar la supervivencia del grupo en un mundo hostil. Una persona baja en este rasgo (poco confiable) sería una elección pobre para proteger el abastecimiento de comida. La capacidad para discernir la escrupulosidad en otros tiene, por tanto, un valor adaptativo. No es de sorprender que los cinco puntos mencionados antes correspondan al modelo de los cinco factores de la personalidad.

Costa y McCrae han desarrollado dos pruebas de personalidad con base en el modelo de los cinco factores (Costa, 1991; McCrae y Costa, 1987). El Inventario NEO de personalidad-Revisado (NEO-PI-R) contiene 240 reactivos calificados en una escala de cinco puntos. Además de los cinco principales dominios de la personalidad, el inventario mide seis rasgos específicos (llamados facetas) dentro de cada uno. También está disponible una versión abreviada con 60 reactivos que se conoce como Inventario NEO de cinco factores (NEO-FFI). Trull, Widiger, Useda y colaboradores (1998) han publicado una entrevista semiestructurada para la evaluación del modelo de los cinco factores de la personalidad. Estas pruebas se analizan en el siguiente capítulo.

COMENTARIO SOBRE EL CONCEPTO DE RASGO

El reto que enfrentan los teóricos de los rasgos es que durante largo tiempo los psicólogos han sabido que en cualquier diccionario estándar del inglés pueden encontrarse miles de nombres de rasgos. Por ejemplo, en uno de los primeros y más influyentes estudios, Allport y Odbert (1936) registraron más de 18 000 nombres de rasgos. Obviamente, éstos son demasiados como para ser útiles en cualquier teoría de personalidad o en una prueba, de modo que es necesario que los teóricos busquen un número más pequeño y manejable de rasgos básicos. Hasta hace poco, no existía un consenso en cuanto al número de rasgos fundamentales. Algunos teóricos propusieron dos o tres factores de rasgo prevalecientes, mientras que otros dividieron el dominio de la personalidad en 16 a 20 dimensiones de rasgo. Muchos teóricos de la personalidad —quizá la mayoría— admiten hoy en día que los cinco factores señalados antes (Neuroticismo, Extraversión, Apertura, Conformidad y Escrupulosidad) proporcionan una manera económica y útil de ver la personalidad. Pero este modelo es muy reciente y requerirá tiempo confirmar su utilidad. Por ejemplo, todavía hay un debate en cuanto a si apertura a la experiencia pertenece a la lista de dimensiones fundamentales de la personalidad (Digman, 1990). También, ¿por qué no se incluye al intelecto dentro del modelo de cinco factores?

Todos los enfoques de rasgos de personalidad comparten ciertos problemas. Primero, existe desacuerdo en cuanto a si éstos causan la conducta o sólo la describen (Fiske, 1986). Puede discutirse de manera persuasiva que invocar los rasgos como causas es una forma vacía de razonamiento circular. Por ejemplo, podría decirse que una persona con normas muy elevadas posee el rasgo de perfeccionismo. Pero cuando se pide explicar qué se quiere decir con perfeccionismo, invariablemente uno termina haciendo referencia a un patrón de normas muy elevadas. Así, cuando se afirma que alguien es perfeccionista, ¿en realidad se está haciendo algo más que proporcionar una descripción abreviada de su conducta pasada? Miller (1991) ha dado voz a esta crítica del enfoque de cinco factores, señalando que el modelo sólo describe la psicopatología, pero no la explica.

Un segundo problema con los rasgos es su validez predictiva aparentemente baja. Se reconoce a Mischel (1968) por haber sido el pri-

mero en desacreditar de manera eficaz el concepto de rasgo en su influyente libro *Personality and Assessment.* Este autor afirmó que "mientras la teoría de los rasgos predice la consistencia conductual, lo que se observa por lo común es la inconsistencia en la conducta" (Mischel, 1968). En una amplia reseña de la investigación existente, Mischel señaló que las escalas de rasgo producían coeficientes de validez con un límite superior de $r = .30$. Dicho autor acuñó el término **coeficiente de personalidad** para describir estas correlaciones bajas. Aunque sin duda son significativas para grandes muestras de individuos, las correlaciones de $r = .30$ son de valor mínimo en la predicción de la conducta individual.

Los investigadores de los rasgos respondieron a los ataques de Mischel refinando y limitando el concepto de rasgo. Éstos buscaron identificar subgrupos de personas cuyas conductas pudieran predecirse de manera precisa con base en las calificaciones de rasgo y también intentaron distinguir los tipos de situaciones en las que éstos determinan en gran medida la conducta (Amelang y Borkenau, 1986; Bem y Allen, 1974; Houts, Cook y Shadish, 1986; Moskowitz, 1982; Wright y Mischel, 1987). Estos esfuerzos lograron un éxito modesto, elevando la validez de algunos cuestionarios de rasgo —en algunos contextos y con ciertas personas— sustancialmente más allá de la ominosa barrera del .30 planteada por Mischel (1968). Pero han quedado lejos los días de las aseveraciones simplistas y generalizadas como la de que "el rasgo X predice la conducta Y".

RESUMEN

1. La personalidad es un constructo vago que se invoca para explicar la consistencia conductual dentro de las personas y la distintividad conductual entre éstas. A fin de apreciar la naturaleza de las pruebas de personalidad, es útil revisar las teorías sobre el tema.
2. Las teorías psicoanalíticas de personalidad se iniciaron con el trabajo original de Sigmund Freud. Según la teoría tripartita de Freud acerca de la mente, la conducta es el resultado dinámico de la lucha entre el ello, yo y superyó. El ello es completamente inconsciente, se orienta al placer y es el asiento de todas las necesidades instintivas, como aquéllas de comida, agua, gratificación sexual y evitación del dolor.
3. Poco después del nacimiento, parte del ello evoluciona para formar el yo o sí mismo consciente. El yo sirve al ello, pero obedece al principio de realidad. El yo también debe enfrentarse al superyó, el componente ético de la personalidad que se modela con base en las normas paternas y sociales de lo que es correcto e incorrecto.
4. Para ayudarse en esta tarea difícil, el yo emplea mecanismos de defensa, que consisten de una variedad de estrategias cognitivas **inconscientes** para protegerse de la ansiedad. Los mecanismos de defensa como la proyección (atribuir las propias faltas a los demás) funcionan porque distorsionan la realidad.
5. En un estudio longitudinal con entrevistas, Vaillant ha mostrado que los mecanismos de defensa tienden hacia mayor madurez en la mitad de la vida. También, el uso de mecanismos de defensa maduros en la adultez joven predicen un mejor resultado en la vida adulta, como se mide a través de criterios independientes como la estabilidad matrimonial, ausencia de problemas con drogas y similares.
6. Las teorías de tipo intentan clasificar a los individuos en categorías o tipos separados. Por ejemplo, el patrón conductual Tipo A, con propensión a la cardiopatía, incluye inseguridad de estatus, hiperagresividad, hostilidad libre flotante y una sensación de urgencia de tiempo (enfermedad de la pri-

sa). Las personas Tipo A —en particular aquéllas con propensión al enojo o a la urgencia de tiempo— pueden estar en mayor riesgo de sufrir cardiopatías e infartos.

7. Las teorías fenomenológicas de la personalidad enfatizan la importancia de la experiencia inmediata, personal y subjetiva como determinante de la conducta. La perspectiva fenomenológica se originó con el filósofo alemán Husserl y con el escritor existencialista danés Kierkegaard.
8. El teórico fenomenológico con mayor influencia fue Carl Rogers, quien creía que el sí mismo y el autoconcepto son centrales para la personalidad. Rogers inventó la clasificación Q para medir el autoconcepto y el sí mismo ideal. En ésta, el individuo examinado clasifica afirmaciones autorreferenciales en aproximadamente nueve pilas ("menos parecido a mí" hasta "más parecido a mí").
9. Una suposición fundamental de todas las teorías conductuales y del aprendizaje social es que muchas de las conductas que forman la personalidad son aprendidas. Los teóricos conductuales radicales, como Skinner, no consideran que las cogniciones tengan algún papel en la explicación de la conducta. En contraste, los teóricos del aprendizaje social, como Rotter, creen que las expectativas (cogniciones) acerca de los reforzadores ambientales son los principales determinantes de la conducta.
10. Guilford define a un rasgo como cualquier manera relativamente duradera en la que un individuo difiere de otro. Las teorías de rasgo surgieron a partir de los modos en que las personas describen a otras en la vida cotidiana. Mischel ha señalado que una de las principales debilidades del enfoque de rasgos es que éstos poseen poca validez predictiva, la cual en pocas ocasiones supera a $r = .30$.
11. La teoría analítico-factorial de Cattell acerca de los rasgos denomina rasgos superficiales a los aspectos más obvios de la personalidad; como la agresividad. Éstos surgen en las primeras etapas del análisis factorial. Los rasgos fuente —más importantes y predictivos de la conducta que los superficiales— se revelan a través del agrupamiento de los rasgos superficiales. El 16FP de Cattell se basa en este modelo.
12. El modelo de los cinco factores propone una síntesis moderna de los enfoques de rasgo en términos de cinco dimensiones de personalidad: Neuroticismo, Extraversión, Apertura, Conformidad y Escrupulosidad. Costa y McCrae han desarrollado dos inventarios basados en este enfoque (NEO-PI-R y NEO-FFI).

TÉRMINOS Y CONCEPTOS CLAVE

Personalidad p. 576
Ello p. 578
Principio del placer p. 578
Yo p. 578
Principio de realidad p. 578
Superyó p. 578
Mecanismos de defensa p. 578
Patrón conductual Tipo A, con propensión a la cardiopatía p. 582
Técnica Q p. 584
Locus de control p. 586
Autoeficacia p. 586
Rasgo p. 587
Rasgos superficiales p. 587
Rasgos fuente p. 588
Hipótesis léxica fundamental p. 588
Coeficiente de personalidad p. 591

TEMA 13B
Técnicas proyectivas

Hipótesis proyectiva
Compendio de técnicas proyectivas
Técnicas de asociación
Técnicas de completamiento
Técnicas de construcción
Técnicas expresivas
Reiteración: La paradoja proyectiva
Ejemplo de caso 13–1. Pruebas proyectivas como auxiliares de la entrevista
Resumen
Términos y conceptos clave

Frank (1939, 1948) introdujo el término **método proyectivo** para describir una categoría de pruebas para el estudio de la personalidad con estímulos desestructurados. En una **prueba proyectiva**, la persona examinada se enfrenta con estímulos vagos, ambiguo y responde con su propias construcciones. Los seguidores de las pruebas proyectivas tienen una fuerte influencia de la teoría psicoanalítica y de su postulado acerca de los aspectos inconscientes de la personalidad. Estos examinadores consideran que los estímulos desestructurados, vagos y ambiguos proporcionan la circunstancia ideal para que se revelen los aspectos internos de la personalidad. La suposición central de las pruebas proyectivas es que las respuestas de una prueba representan proyecciones de los procesos mentales inconscientes más internos de la persona examinada. Se presenta este tema con algunos conceptos y distinciones preliminares que conciernen a las pruebas proyectivas.

HIPÓTESIS PROYECTIVA

La suposición de que las interpretaciones personales acerca de los estímulos ambiguos deben reflejar por obligación las necesidades, motivos y conflictos inconscientes del individuo examinado se conoce como **hipótesis proyectiva**. En general, Frank (1939) recibe el crédito de haber popularizado la hipótesis proyectiva:

> Cuando se exploran los procedimientos existentes que pueden llamarse métodos proyectivos, se encuentra una amplia variedad de técnicas y materiales que se emplean para el mismo propósito general, obtener del individuo "aquello que no puede o no quiere decir", con frecuencia porque no lo sabe y no está consciente de lo que revela sobre sí mismo a través de sus proyecciones.

El reto de las pruebas proyectivas consiste en descifrar los procesos de personalidad subyacentes (necesidades, motivos y conflictos) con base en las respuestas individualizadas, únicas y subjetivas de cada persona examinada. En las secciones se analizará qué tan bien satisfacen las pruebas proyectivas esta tarea portentosa.

COMPENDIO DE TÉCNICAS PROYECTIVAS

ORÍGENES DE LAS TÉCNICAS PROYECTIVAS

Las técnicas proyectivas datan del siglo pasado. Con el propósito de una revisión rápida, se debe decir que Galton (1879) desarrolló la primera técnica proyectiva, una prueba de asociación de palabras. Kent y Rosanoff (1910) adaptaron este procedimiento a las pruebas y C. G. Jung y otros las utilizaron en terapia. Mientras tanto, Ebbinghaus (1897) utilizó una prueba de oraciones incompletas como medida de la inteligencia, pero pronto otros investigadores se dieron cuenta de que el método se adaptaba mejor a la evaluación de la personalidad (Payne, 1928; Tendler, 1930). Con una fuerte influencia de las formulaciones psicoanalíticas acerca de la personalidad, Rorschach publicó su famosa prueba de manchas de tinta en 1921. En 1905, Binet inventó un precursor de las técnicas de narración de historias o de apercepción temática cuando utilizó respuestas verbales hacia ilustraciones como medida de la inteligencia. Éstos y otros intentos forman la piedra angular de las pruebas proyectivas actuales.

LA POPULARIDAD DE LAS PRUEBAS PROYECTIVAS: UNA PARADOJA

El extenso uso de las pruebas proyectivas ha continuado sin disminuir desde principios del siglo XX hasta nuestros días (Louttit y Browne, 1947; Lubin, Wallis y Paine, 1971: Watkins, Campbell y McGregor 1988). De manera reciente, Watkins, Campbell, Nieberding y Hallmark (1995) sometieron a una encuesta a más de 400 psicólogos que practicaban la evaluación para estimar la frecuencia de utilización de diversas pruebas destacadas. Descubrieron que 5 de las 15 pruebas de uso más frecuente eran técnicas proyectivas (cuadro 13–3).

Paradójicamente, desde el punto de vista de los criterios psicométricos tradicionales, las pruebas proyectivas no resultan tan bien libradas como las objetivas que se analizan en el siguiente capítulo. El enigma esencial de las pruebas proyectivas es cómo explicar la duradera popularidad de estos instrumentos a pesar de su calidad psicométrica algunas veces cuestionable. Después de todo, no todos los psicólogos son insensatos, como tampoco ignorantes de los problemas sobre calidad de una prueba. Entonces, ¿por qué persisten las técnicas proyectivas? Se regresará a este enigma —que podría llamarse la paradoja proyectiva— después de que se familiarice al lector con los principales enfoques dentro de las pruebas proyectivas.

UNA CLASIFICACIÓN DE LAS TÉCNICAS PROYECTIVAS

Lindzey (1959) ha ofrecido una clasificación de las técnicas proyectivas que se seguirá aquí.

Figura 13–3. Las 15 pruebas de utilización más frecuente en EUA

Prueba	Rango
Escala Wechsler de Inteligencia para Adultos—Revisada	1
Inventario Multifásico de la Personalidad—2	2
Métodos de frases incompletas.*	3
Test de Apercepción Temática.*	4
Rorschach	5
Test Guestáltico Visomotor	6
Dibujos proyectivos.*	7
Beck Depression Inventory (Inventario de Depresión Beck)	8
Escala Wechsler de Inteligencia para Niños—III	9
Wide Range Achievement Test—Revised (Prueba de Rendimiento de Amplio Rango —Revisada)	10
Wechsler Memory Scale—Revised	11
Peabody Picture Vocabulary Test—Revised	12
Millon Clinical Multiaxial Inventory —II (Inventario Clínico Multiaxial de Millon—II)	13
Escala para los Niveles Preescolar y Primario—R	14
Test de Apercepción Infantil.*	15

* Denota una prueba proyectiva. Algunos examinadores utilizan el Test Guestáltico Visomotor como prueba proyectiva.

Fuente: Adaptado con autorización de Watkins, C., Campbell, V., Nieberding, R. y Hallmark, R. (1995). Contemporary practice of psychological assessment by clinical psychologists. *Professional Psychology; Research and Practice, 26*, 54-60.

Con base en las respuestas requeridas, dividió las pruebas proyectivas en cinco categorías:

- Asociación con manchas de tinta o palabras.
- Construcción de historias o secuencias.
- Completamiento de frases o historias.
- Ordenación/selección de ilustraciones u opciones verbales.
- Expresión con dibujos o juegos.

Las técnicas de asociación incluyen la ampliamente utilizada prueba de interpretación de Formas de Rorschach y su prima, la *Holtzman Inkblot Test* (Prueba Holtzman de manchas de tinta), la cual se considera psicométricamente superior, al igual que las pruebas de asociación de palabras. Las técnicas de construcción incluyen el Test de apercepción temática y muchas variaciones de este primer instrumento. Las técnicas de completamiento consisten principalmente de las pruebas de Frases incompletas, que se analizan adelante. Los procedimientos de ordenación/selección como la prueba Szondi (la cual se analiza en el primer capítulo) se utilizan rara vez en la actualidad. Por último, las técnicas expresivas, como el Dibujo de la figura humana o la Técnica del dibujo proyectivo de Casa-árbol-persona son muy populares entre los clínicos, a pesar de sus dudosos datos de validez.

Se revisarán las técnicas más notables dentro de cada categoría, excepto los anticuados enfoques de ordenación/selección que casi nunca se emplean. No obstante, la literatura sobre las principales técnicas proyectivas es simplemente abrumadora, con decenas de miles de artículos nada más acerca del Rorschach. Se pueden sugerir las principales tendencias en cuanto a investigación, pero será necesario que el lector consulte otras fuentes para obtener una reseña amplia.

TÉCNICAS DE ASOCIACIÓN

RORSCHACH

El Rorschach consiste en 10 manchas de tinta diseñadas por Herman Rorschach (1884-1922) a inicios del siglo XX; formó las manchas al derramar tinta en una hoja de papel y doblarla por la mitad, produciendo diseños bilaterales relativamente simétricos. Cinco de las manchas son negras o con sombreado gris, mientras cinco contienen color; cada una se presenta sobre un fondo blanco. En la figura 13–2 se muestra una mancha de tinta del tipo de las empleadas



Figura 13–2. Una mancha de tinta similar a las encontradas en el Rorschach.

por Rorschach. Esta prueba es adecuada para personas de cinco años de edad en adelante, pero por lo común se emplea con adultos.

Para la aplicación del Rorschach, el examinador se sienta al lado de la persona para reducir al mínimo la comunicación a través del lenguaje corporal. La aplicación consiste de dos fases. En la fase de asociación libre, el examinador presenta la mancha de tinta y pregunta "¿Qué podría ser esto?" Si la persona examinada pide aclaración (p. ej., "¿Debo utilizar la mancha completa o sólo parte de ella?"), el examinador siempre responde de modo no directivo ("Como prefiera"). La prueba procede de manera pausada, de modo que existe una expectativa implícita de que la persona dará más de una respuesta por lámina. No obstante, no se requiere esto; incluso es permisible que el individuo rechace la lámina por completo, aunque esto sucede rara vez. Las 10 láminas se presentan de modo similar.

A continuación el examinador comienza con la fase de encuesta. En ésta el examinador hace preguntas para aclarar la localización exacta en la mancha de cada elemento percibido y para determinar cuáles aspectos de la misma, como forma o color, representaron un papel en la creación de la respuesta. Con base en la información recabada durante la fase de encuesta, el examinador puede codificar la localización, determinantes, calidad de la forma y contenido de cada respuesta de acuerdo con uno o más sistemas formales de calificación. Por ejemplo, si la persona evaluada utilizó la mancha completa para su percepto, la respuesta se codifica como *W* (completa); si la forma de la mancha fue importante, la respuesta también se codifica como *F* (forma); si en el percepto aparece movimiento humano, la respuesta se codifica como *M* (movimiento); el uso del color en una respuesta se codifica con *C* (color), *CF* (color/forma) o *FC* (forma/color), dependiendo de si la forma está totalmente ausente, es primaria o secundaria al color como determinante. También se codifica el contenido del percepto, por ejemplo H (humano), Hd (detalle humano), An (anatomía), Cg (vestimenta) y demás (cuadro 13–4). La calificación apropiada del Rorschach requiere entrenamiento y supervisión extensas; aquí sólo se han tocado unos cuantos aspectos básicos.

Es lamentable que Rorschach haya muerto antes de completar sus métodos de calificación, de modo que la sistematización de la calificación de la prueba le quedó a sus seguidores. Cinco psicólogos estadounidenses produjeron enfoques superpuestos, pero independientes, para la prueba —Samuel Beck, Marguerite Hertz, Bruno Klopfer, Zygmund Piotrowski y David Rapaport (Erdberg, 1985). Como podría predecirse, los matices de la calificación varían de un método a otro. Por fortuna, Exner y sus colaboradores han sintetizado estos enfoques anteriores dentro de un Sistema comprensivo de calificación (Exner, 1991, 1993; Exner y Weiner, 1994). Este sistema se fundamenta mejor en la investigación empírica y ha reemplazado con toda claridad a los demás enfoques para la calificación de la prueba.

Una vez que se ha codificado el protocolo completo, el examinador puede calcular varias calificaciones de resumen que forman la primera base para las hipótesis sobre la personali-

Cuadro 13–4. Resumen de los principales criterios de calificación del Rorschach

I.	Localización: ¿En qué parte de la mancha se localizó el percepto?		
	W	Completa	Se utilizó toda la mancha de tinta
	D	Detalle común	Se utilizó una parte bien definida
	Dd	Detalle inusual	Se utilizó una parte inusual
	S	Espacio	Percepto definido por espacio en blanco
II.	Determinante: ¿Qué característica de la mancha determinó la respuesta?		
	F	Forma	Aspecto o perfil utilizado
	F+	Forma +	Excelente equiparación entre percepto y mancha de tinta
	F–	Forma –	Equiparación muy deficiente entre percepto y mancha de tinta
	M	Movimiento	Movimiento visto o implícito en el percepto
	C	Color	El color ayudó a determinar la respuesta
	T	Textura	El sombreado se incluye en la respuesta
III.	Contenido: ¿Qué es el percepto?		
	H	Humano	Percepto de forma humana completa
	Hd	Detalle humano	Forma humana incompleta en algún sentido
	Ex	Explosión	Una explosión en sí
	Xy	Rayos X	Radiografía de cualquier parte humana; implica sombreado
IV.	Popular contra original		
	P	Popular	Respuesta dada por muchas personas normales
	O	Original	Respuesta rara y creativa

Nota: Este cuadro representa un consenso de todos los principales sistemas de calificación. La lista es incompleta y sólo ilustrativa. Los sistemas completos de calificación son muy complejos y permiten combinaciones; por ejemplo, FM, CF, WS-, Do. Para mayores ejemplos véase Exner, J. E., Jr. (1993). *The Rorschach: A comprehensive system, Volume I. Basic foundations* (3a. ed.). Nueva York: Wiley.

dad del individuo evaluado. Por ejemplo, el porcentaje de F+ es la proporción del total de respuestas que utilizan la forma pura como determinante. Existe una literatura voluminosa acerca del significado de este índice, pero parece seguro hipotetizar que cuando el porcentaje de F+ cae por debajo de 70%, el examinador debería considerar la posibilidad de psicopatología, organicidad o déficit intelectual graves en la persona examinada (Exner, 1993). El porcentaje de F+ se considera también como un índice de la fortaleza del yo, donde las calificaciones más altas indican mayor capacidad para enfrentarse de manera eficaz al estrés. No obstante, en el mejor de los casos se puede decir que existe evidencia a favor y en contra de esta conjetura.

Frank (1990) ha enfatizado que la calificación formal del Rorschach es insuficiente para algunos propósitos, como el diagnóstico de la esquizofrenia. Este autor destaca que un análisis del pensamiento del paciente en busca de la presencia de asociaciones sumamente personales, ilógicas y extrañas hacia las manchas de tinta es esencial para el psicodiagnóstico. Desde su enfoque, el Rorschach es en realidad un auxiliar de la entrevista y no una prueba en sí.

COMENTARIO SOBRE EL RORSCHACH

Debido a una variedad de razones, es difícil ofrecer generalizaciones concisas acerca de la confiabilidad, validez y utilidad clínica del

Rorschach. Incluso las preguntas sencillas provocan respuestas complejas. Por ejemplo, ¿cuál es el propósito de una evaluación con el Rorschach? En épocas sucesivas de investigación, esta prueba se ha utilizado para obtener un diagnóstico psiquiátrico, estimar el pronóstico de la psicoterapia, obtener un índice del proceso primario de pensamiento, predecir el suicidio y formular estructuras complejas de personalidad, por nombrar sólo unas cuantas aplicaciones (Peterson, 1978). El propósito del Rorschach está tan mal definido que inclusive algunos de sus partidarios evitan considerarlo como una prueba y en vez de ello prefieren denominarlo un método para generar información acerca del funcionamiento de la personalidad (Weiner, 1994). Cuando el propósito de un instrumento es poco claro, la investigación objetiva sobre sus atributos psicométricos es arriesgada tanto como difícil. Peor aún, la investigación objetiva puede carecer de sentido, dado que aquellos que lo apoyan ignorarán las conclusiones contrarias, mientras que los detractores no utilizan la prueba de todos modos.

Un estudio de Albert, Fox y Kahn (1980) acerca de la susceptibilidad del Rorschach a la falsificación, es típico de la investigación sobre este instrumento. Se recuerda al lector que existen miles de estudios de investigación dentro de la literatura, incluyendo muchos con conclusiones positivas (p. ej., Hilsenroth, Fowler, Padawer y Handler, 1997; Smith, Gacono y Kaufman, 1997; Weiner, 1996). Pero los resultados heterogéneos que presentaron Albert, Fox y Kahn (1980) son más comunes. Estos investigadores sometieron los protocolos del Rorschach de 24 personas a un panel de expertos, a quienes pidieron sus diagnósticos psiquiátricos para cada individuo. Los 24 protocolos del Rorschach consistieron de los resultados de cuatro grupos con seis personas cada uno:

- Pacientes de un hospital para enfermedades mentales con un diagnóstico de esquizofrenia paranoide.
- Falsificadores no informados a quienes se dio la instrucción de falsear las respuestas de un esquizofrénico paranoide.
- Falsificadores informados que escucharon una cinta de audio detallada acerca de la esquizofrenia paranoide.
- Controles normales que realizaron la prueba bajo condiciones estándar.

Los falsificadores no informados, los informados y los controles normales eran estudiantes que habían pasado por una prueba de detección con el MMPI y a quienes, durante una entrevista, se les había considerado como razonablemente normales. De 6 a 9 jueces, todos miembros de la *Society for Personality Assessment*, calificaron los protocolos. A éstos se les dijo que proporcionaran un diagnóstico psiquiátrico así como otra información que no se proporciona aquí. No se les informó el propósito del estudio, pero se les dijo que evaluaran si parecía haber simulación en cualquiera de los perfiles.

Los falsificadores informados debieron haber hecho un excelente trabajo porque fue más probable que se les diagnosticara como psicóticos que a los pacientes reales en sí (72% contra 48%, respectivamente). Los falsificadores no informados fueron también bastante convincentes, con una tasa de 46% de diagnóstico de psicosis. Los controles normales se diagnosticaron como psicóticos en 24%. Se concede que el reto diagnóstico de este estudio era inmenso, pero de todos modos sigue siendo perturbador encontrar que jueces expertos calificaron a 24% de los protocolos normales como psicóticos, mientras que identificaron psicosis en sólo 48% de los protocolos psicóticos reales. Un estudio más reciente de Netter y Viglione (1994) también concluyó que el Rorschach es susceptible a la simulación de psicosis.

Aunque existen excepciones dignas de mención en la evaluación del Rorschach, un sustancial número de estudios apunta a una baja confiabilidad y a una falta general de validez predictiva (Carlson, Kula y St. Laurent, 1997; Peterson, 1978; Lanyon, 1984; Wood, Nezworski y Stejskal, 1996). En una reseña metaanalítica, Garb, Florio y Grove (1998) concluyeron que el Rorschach explicaba un atroz 8 a 13% de la varianza en las características del cliente, en comparacion con el MMPI, el cual explicaba

de 23 a 30% de la varianza. Desde una perspectiva positiva, estudios recientes basados en las mejorías en calificación que ofrece el enfoque de Exner son un tanto más positivas en sus resultados (véase Exner, 1995; Exner y Andronikof-Sanglade, 1992; Meyer, 1997; Ornberg y Zalewski, 1994; Piotrowski, 1996). Sin embargo, el Rorschach todavía no obtiene el nivel de respetabilidad científica que disfrutan otras pruebas de personalidad y quizá nunca lo logre.

TÉCNICA HOLTZMAN DE MANCHAS DE TINTA

Wayne H. Holtzman buscó superar las principales limitaciones del Rorschach al desarrollar una técnica completamente nueva que utiliza manchas de tinta con procedimientos simplificados para su aplicación y calificación. En la *Holtzman Inkblot Technique*, HIT (Técnica Holtzman de Manchas de Tinta), la persona examinada se limita a una respuesta por lámina, pero se le presenta una serie de 45 láminas. A cada respuesta le sigue una pregunta dual muy sencilla: ¿dónde se encuentra representado el percepto dentro de la mancha y qué aspecto de ésta lo sugirió?

La HIT puede obtenerse en dos formas paralelas, construidas de manera cuidadosa. La existencia de formas paralelas es invaluable para los estudios *test-retest*, ya que es frecuente que las personas examinadas recuerden sus respuestas dadas a una lámina y, por tanto, ofrezcan la misma respuesta de manera mecánica cuando se les vuelve a examinar. Las 45 respuestas de la HIT se califican en cuanto a 22 variables diferentes, obtenidas de los anteriores sistemas de calificación del Rorschach. Las variables de calificación de la HIT se describen en el cuadro 13–5.

El sistema de calificación de la HIT es sumamente confiable y la estandarización del instrumento parece adecuada. Cuando se emplea a calificadores bien entrenados, la concordancia intercalificadores para las diferentes categorías es de .95 a 1.00 para la mayor parte de ellas; sólo Penetración e Integración caen por debajo de esta norma. Las confiabilidades por mitades también son aceptables, con valores medianos en los rangos de .70 y .80. La estabilidad *test-retest* con formas paralelas es adecuada en términos generales, aunque algunas categorías (Localización, con r de .81) tienen mejor desempeño que otras (Popular, con r de .36). Las normas percentilares para cada categoría de calificación se presentan por separado para estudiantes universitarios (N = 206), adultos promedio (N = 252), niños de secundaria (N = 197), niños de escuela primaria (N = 132), niños de cinco años de edad (N = 122), esquizofrénicos crónicos (N = 140), pacientes con depresión (N = 90) y personas con retraso mental (N = 100).

Varios cientos de estudios de investigación que se han dirigido a la validez de la HIT informan sobre la relación entre las calificaciones de la prueba y medidas independientes de personalidad (Hill, 1972; Holtzman, 1988; Swartz, Reinehr y Holtzman, 1983). En general, las relaciones son modestas, pero sustentan la validez de la HIT, en especial como auxiliar para el psicodiagnóstico.

Una variación reciente de la HIT requiere dos respuestas para cada una de las 25 láminas de un subconjunto cuidadosamente seleccionado de la Forma A. Llamada la HIT 25 para distinguirla de la HIT estándar, esta nueva prueba constituye una promesa excepcional para ayudar al diagnóstico de la esquizofrenia. Al utilizar criterios de calificación de gran objetividad y reglas simples de decisión, la HIT 25 clasificó de manera objetiva a 26 de 30 esquizofrénicos y a 28 de 30 estudiantes universitarios normales (Holtzman, 1988). Los criterios de decisión consisten de cuatro reglas para conclusiones normales, que se califican con +1 cada una, y 13 reglas para conclusiones de esquizofrenia, que se califican como –1 cada una. Los resultados totales se suman de manera algebraica, lo cual produce una calificación de "normalidad". Ésta es la base para decisiones diagnósticas simples. Las calificaciones por arriba de cero sugieren normalidad, mientras que las inferiores a cero indican esquizofrenia: una calificación de cero es indeterminada. La HIT 25 parece prometedora, pero se necesitan especialmente estudios sobre validación cruzada.

Cuadro 13–5. Nombres y descripciones de las variables en Técnica Holtzman de manchas de tinta

Tiempo de reacción:	Tiempo en segundos desde la presentación de la mancha de tinta hasta el inicio de la primera respuesta.
Rechazo:	El individuo no informa nada o regresa la mancha de tinta al examinador.
Localización:	Calificada en un sistema de 3 puntos: 0 –mancha completa, 1 –área grande, 2 –área pequeña.
Espacio:	Calificado cuando existe una verdadera inversión figura-fondo: la parte en blanco constituye la figura.
Definición de forma:	Calificada en un sistema de 5 puntos desde 0 (concepto sin forma, p. ej., mancha de pintura) hasta 4 (concepto sumamente formado, p. ej., centauro).
Idoneidad de forma:	Bondad de ajuste del concepto a la forma de la mancha de tinta; 0 –deficiente, 1 –adecuada, 2 –buena.
Color:	El color es el principal determinante, en general lo menciona el sujeto; se califica de 0 a 3.
Sombreado:	El individuo hace referencia al sombreado (borroso, con textura) como un determinante; se califica de 0 a 2.
Movimiento:	Se califica cuando la respuesta implica energía o una cualidad de movimiento dinámico; se califica de 0 a 4.
Verbalización patognomónica:	Verbalizaciones incoherentes, raras, absurdas, autorreferenciales, etc., ante las láminas.
Integración:	Calificada con 1 si dos o más elementos de la mancha se integran efectivamente en la respuesta; de otra manera se califica con 0.
Calificaciones de contenido:	Cada categoría (Humano, Animal, Anatomía, Sexo, Abstracto) se califica de 0 a 2 con base en ausencia, presencia parcial o presencia completa del concepto.
Ansiedad:	Cada respuesta se califica de 0 a 2 en cuanto a signos de ansiedad (p. ej., cueva oscura y peligrosa).
Hostilidad:	Cada respuesta se califica de 0 a 3 en cuanto a signos de hostilidad (p. ej., mariposa destrozada).
Barrera:	La barrera se refiere a cualquier cobertura, membrana, concha o piel protectora que pudiera relacionarse a nivel simbólico con límites de imagen corporal; 1 si está presente, 0 si está ausente.
Penetración:	Se califica con 1 si el concepto es símbolo de que el individuo examinado siente que algo puede penetrar con facilidad su cuerpo exterior; de otro modo se califica con 0.
Equilibrio:	Se califica con 1 si la persona hace referencia a la presencia o ausencia de simetría en el diseño; de otro modo se califica con 0.
Popular:	Se califica con 1 si la respuesta es común, observada en 1 de 7 protocolos normativos.

Fuente: Basado en Holtzman, W. H., (1961). *Guide to administration and scoring: Holtzman Inkblot Technique.* New York: The Psychological Corporation.

TÉCNICAS DE COMPLETAMIENTO

PRUEBAS DE FRASES INCOMPLETAS

En una prueba de frases incompletas, se presenta a la persona examinada una serie de partes de frases que se forman de unas cuantas palabras del inicio de la misma y la tarea consiste en dar las terminaciones. Como con cualquier técnica proyectiva, el examinador supone que las oraciones terminadas reflejan las motivaciones, actitudes, conflictos y temores subyacentes de la persona que responde a la prueba. En general, las pruebas de frases incompletas se pueden interpretar de dos maneras diferentes: análisis subjetivo-intuitivo de las motivaciones subyacentes proyectadas en las respuestas del individuo o análisis objetivo por medio de calificaciones asignadas a cada oración terminada.

En la figura 13–3 se presenta un ejemplo de una prueba de completamiento de frases. Ésta es bastante similar a los instrumentos existentes en cuanto a que las partes de la frase son muy cortas y se restringen a un pequeño número de temas básicos. El lector observará que en esta prueba corta se repiten tres temas (autoconcepto, la madre y el padre de la persona que responde). De esta manera, el individuo examinado tiene múltiples oportunidades de revelar sus motivaciones subyacentes acerca de cada tema. Por supuesto, la mayor parte de las pruebas de frases incompletas son más grandes —entre 40 y 100 partes de una frase— y contienen más temas —de 4 a 15 temas.

Se han desarrollado docenas de pruebas de frases incompletas, la mayor parte de las cuales son instrumentos no publicados y sin estandarizar que se produjeron para satisfacer una necesidad clínica específica. En el cuadro 13–6 se describen algunas de las pruebas de completamiento de frases más destacadas y que todavía están en uso. De estos instrumentos, la Prueba de la Universidad de Washington de Frases Incompletas, creada por Loevinger, es la más compleja y teóricamente fundamentada (p. ej., Weiss, Zilberg y Genevro, 1989). No obstante, el *Rotter Incomplete Sentences Blank* (Formulario de frases incompletas de Rotter) tiene la sustentación empírica más fuerte y es la de más amplia utilización en ambientes clínicos. Se analiza en más detalle este instrumento.

FORMULARIO DE FRASES INCOMPLETAS DE ROTTER

El *Rotter Incomplete Sentences Blank* (RISB) consiste en tres formas similares: educación media, universitaria y adulta; cada una de las cuales contiene 40 partes de frases escritas en su mayor parte en primera persona (Rotter y Rafferty, 1950; Rotter, Lah y Rafferty, 1992). Aunque la prueba puede interpretarse subjetivamente de la manera común, a través de un

Instrucciones: Termine estas frases para indicar cómo se siente usted.

1. Mi mejor característica es _____
2. Mi madre _____
3. Mi padre _____
4. Mi mayor temor es _____
5. Lo mejor de mi madre era _____
6. Lo mejor de mi padre era _____
7. Estoy orgulloso(a) de _____
8. Sólo quisiera que mi madre hubiera _____
9. Sólo quisiera que mi padre hubiera _____

Figura 13–3. Ejemplo de una prueba corta de frases incompletas.

Cuadro 13–6. Resumen de pruebas importantes de frases incompletas

Bloom Sentence Completion Survey* (BSCS) (1974-1975) (Examen de Frases Incompletas de Bloom), *Stoelting Co.

El BSCS puede obtenerse en una forma con dos niveles (estudiantes o adultos). Cuarenta partes de frases producen calificaciones en ocho áreas: compañeros por edad (estudiantes) o personas (adultos), sí mismo físico, familia, sí mismo psicológico, independencia, educación (estudiantes) y trabajo (adultos), logros e irritantes.

Activity Completion Technique* (1984) (Técnica de Completamiento de Actividad), *Psychological Assessment Resources

La ACT consiste de 60 partes de frases y cubre cuatro áreas: Familia, Interpersonal, Afecto y Autoconcepto. El manual revisa un análisis de contenido de las respuestas con orientación clínica dentro de cada área. La prueba también emplea un sistema semiobjetivo de calificación para producir una Escala de calificación de adaptación.

Incomplete Sentences Task* (IST) (1979-1980) (Tarea de Frases Incompletas), *Stoelting Co.

La IST consiste de dos formas (niveles bachillerato y universidad) con 39 reactivos sumamente estructurados, vinculados a la conducta, que se diseñaron en especial para evaluar hostilidad, ansiedad y dependencia.

Geriatric Sentence Completion Form* (1982) (Forma geriátrica de Frases Incompletas), *Psychological Assessment Resources

La GSCF es una forma con 30 reactivos diseñada específicamente para utilizarse con clientes adultos mayores. Esta prueba obtiene respuestas personales dentro de cuatro dominios de contenido: físico, psicológico, social y de orientación temporal. El manual de la prueba incluye varios ejemplos clínicos.

***Washington University Sentence Completion Test* (Prueba de oraciones incompletas de la Universidad de Washington), publicada de manera privada por Loevinger (1976, 1979)**

La WUSC utiliza formas independientes para varones, mujeres y varones y mujeres más jóvenes. Esta prueba tiene una vinculación muy fuerte con la teoría; las respuestas se clasifican según siete etapas de desarrollo del yo; presocial y simbiótica, impulsiva, autoprotectora, conformista, escrupulosa, autónoma, íntegra.

Miner Sentence Completion Scale* (Escala Miner de oraciones incompletas) (Miner, 1964, 1986), *Organizational Measurement Systems Press

La MSCS consiste de tres formas, incluyendo la Forma *H* original que se publicó inicialmente en 1964. La prueba se dirige a empleos administrativos, profesionales y orientados a las ventas. La Forma *H* es única en proporcionar una alternativa de opción múltiple; las nueve subescalas incluyen dimensiones relacionadas con el trabajo como Figuras de autoridad, Situaciones competitivas y Rol asertivo.

análisis cualitativo de las necesidades proyectadas en las respuestas del individuo, es la calificación objetiva y cuantitativa del RISB la que ha atraído mayor atención.

En el sistema objetivo de calificación, cada una de las frases completadas recibe una calificación de adaptación de 0 (buena adaptación) a 6 (adaptación muy deficiente). Éstas se basan de inicio en la siguiente categorización de cada respuesta:

- **Omisión** –no hay respuesta o ésta es demasiado corta como para tener significado.
- **Respuesta de conflicto** –indicativa de hostilidad o infelicidad.
- **Respuesta positiva** –indicativa de actitud positiva o esperanzada.
- **Respuesta neutra** –afirmaciones declarativas que no demuestran afecto positivo ni negativo.

Ejemplos de las últimas tres categorías incluyen:

Odio... a todo el mundo (respuesta de conflicto).
Lo mejor... aún está por venir (respuesta positiva).
La mayoría de las niñas... son mujeres (respuesta neutra).

Las respuestas de conflicto se califican con 4, 5 o 6, del menor al mayor grado de conflicto expresado. Las respuestas positivas se califican como 2, 1 o 0, de la menos positiva a la más positiva. El manual proporciona ejemplos de cada categoría de calificación. La calificación total de adaptación se obtiene al sumar las calificaciones ponderadas en las categorías de conflicto y positiva. La calificación de adaptación puede variar de 0 a 240, donde las calificaciones más altas indican mayor desadaptación.

La confiabilidad de la calificación de adaptación es excepcionalmente buena, incluso cuando la obtienen asistentes con conocimientos psicológicos mínimos. Por lo común, las confiabilidades intercalificadores se encuentran en el rango de .90 y los coeficientes por mitades están en el rango de .80 (Rotter, Lah y Rafferty, 1992; Rotter, Rafferty y Schachtitz, 1965). La validez de este índice se ha investigado en numerosos estudios que emplean el RISB como instrumento de detección con una puntuación de corte para la "desadaptación". Por ejemplo, se ha encontrado que una puntuación de corte de 135 detecta de manera correcta a los delincuentes juveniles en 60% de las ocasiones, mientras que también identifica correctamente a los jóvenes no delincuentes en 73% de las veces (Fuller, Parmelee y Carroll, 1982). Las mismas puntuaciones de corte identifican a los individuos con fuerte abuso de drogas en 80 a 100% de las ocasiones (Gardner, 1967). Éstas y otras conclusiones similares sustentan la validez de constructo del índice de adaptación, pero también indican que las tasas de clasificación son mucho más bajas que las necesarias para la toma de decisiones individuales o la detección eficaz. También parece que las normas del índice de adaptación son anticuadas. Lah y Rotter (1981) encontraron que las calificaciones actuales de los estudiantes difieren de manera significativa de las obtenidas en el estudio original de Rotter y Rafferty (1950). Lah (1989) y Rotter, Lah y Rafferty (1992) proporcionan nuevos datos normativos, de calificación y de validez para el RISB.

Como propone Goldberg (1965), la sencillez de la calificación única de adaptación constituye tanto la fortaleza como la debilidad de la prueba. Es cierto, el instrumento proporciona un método rápido y eficaz para obtener un índice general del funcionamiento cotidiano de las personas. No obstante, una sola calificación no tiene posibilidades de captar los matices del funcionamiento de la personalidad. Además, el RISB está sujeto al mismo tipo de sesgo que otras medidas de autoinforme; es decir, la información reflejará principalmente aquello que la persona que responde quiere que el examinador sepa (Phares, 1985).

ESTUDIO ROSENZWEIG DE FRUSTRACIÓN POR ILUSTRACIONES

El *Rosenzweig Picture Frustration Study* (Estudio Rosenzweig de Frustración por Ilustraciones; Estudio P-F) se considera con frecuencia como una técnica semiproyectiva que requiere que el examinando produzca una respuesta verbal a un estímulo verbal-pictórico sumamente estructurado. El Estudio P-F se puede obtener en tres formas —niños, adolescentes y adultos— cada una de las cuales consiste de 24 ilustraciones de tira cómica que representan una circunstancia frustrante (Rosenzweig, 1977, 1978a). Cada lámina contiene dos personas, donde la que se encuentra a la izquierda expresa palabras que provocan o describen una situación frustrante para la persona a la derecha (figura 13–4). Al individuo examinado se le pide que indique, al escribir en el globo que se encuentra arriba de la cabeza de la persona frustrada, la primera respuesta que le venga a la mente como expresada por la figura anónima de la caricatura. En el caso de los individuos más jóvenes, el examinador escribe la respuesta de la persona.

El propósito del Estudio P-F consiste en evaluar la manera característica en que la persona examinada reacciona ante la **frustración**. Ésta se define como aquello que ocurre cuando el organismo se enfrenta con un obstáculo u obstrucción en su camino a la satisfacción de una necesidad (Rosenzweig, 1944). En sentido general, es bien sabido que las personas reac-

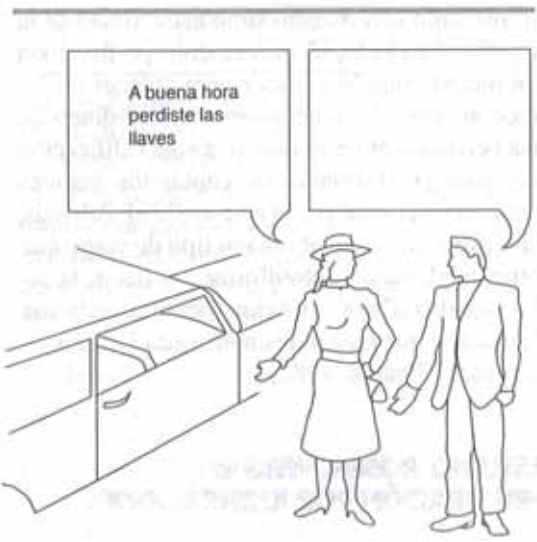


Figura 13–4. Reactivo de muestra del Estudio Rosenzweig de Frustración por Ilustraciones.
Copyright © de Saul Rosenzweig. Reproducido con autorización.

cionan a la frustración con agresión. El valor del Estudio P-F es su conceptuación multifacética de la agresión, según tres direcciones y tres tipos. La dirección de la agresión puede ser extragresiva; es decir, volcada hacia el ambiente; intraagresiva (o sea, volcada por la persona examinada hacia sí mismo), o imagresiva o evitada en un intento por disimular la frustración. El tipo de agresión puede ser con dominio del obstáculo, en el cual se destaca en la respuesta la barrera que ocasiona la frustración; en defensa del yo, en la cual la capacidad de organización de la persona examinada predomina en la respuesta; o de persistencia de la necesidad, en la que se enfatiza la solución de la situación frustrante al seguir hacia la meta a pesar del obstáculo (Rosenzweig, 1978b). Es importante señalar que la agresión no necesariamente es un constructo negativo. Los tipos de agresión de persistencia de la necesidad representan formas constructivas, y en ocasiones creativas, de agresión, mientras que la agresión en defensa del yo es con frecuencia destructiva (de otros o de uno mismo).

El Estudio P-F se califica al de detectar uno o dos de los factores en cada respuesta individual. Se evitan las interpretaciones profundas y el manual contiene muestras de calificación para ayudar en la toma de decisiones. Cuando se han contado las calificaciones de los reactivos, se llena el formulario de calificación calculando el porcentaje de las nueve categorías de calificación que ocurrieron en el protocolo de la persona evaluada. También se cuentan los tipos y direcciones generales de la agresión, lo cual da por resultado 15 índices. Además se puede calcular una Calificación de conformidad con el grupo (GCR). La GCR indica el grado de correspondencia entre las respuestas del individuo examinado y aquéllas dadas con más frecuencia por la muestra normativa. Todos los índices pueden compararse con los resultados de muestras apropiadas de estandarización. Por supuesto, además de la calificación cuantitativa, las respuestas al Estudio P-F pueden evaluarse de manera impresionista.

La confiabilidad intercalificadores del Estudio P-F se presenta en el rango de .80 a .85 para examinadores bien entrenados y escrupulosos. Sin embargo, la estabilidad *test-retest* del instrumento se encuentra en algún punto entre promedio y marginal. Por ejemplo, las correlaciones *retest* para las categorías de calificación en la forma adulta del Estudio P-F van de .21 a .71, donde la mayor parte de los valores se colocan en el rango de .40 (Rosenzweig, 1978b). Una enorme cantidad de estudios de validación se encuentran resumidos en varias publicaciones (Rosenzweig, 1977, 1978b; Rosenzweig y Adelman, 1977). Con base en las confiabilidades muy modestas de las categorías de calificación, se coincide en que el Estudio P-F es más apropiado para la investigación que para la evaluación individual (Graybill y Heuvelman, 1993).

TÉCNICAS DE CONSTRUCCIÓN

TEST DE APERCEPCIÓN TEMÁTICA (TAT)

La TAT consiste de 30 ilustraciones que representan una variedad de cuestiones y temas en dibujos y fotografías en blanco y negro; una de

las láminas está en blanco. La mayor parte de las láminas representan a una o más personas que participan en actividades ambiguas. Algunas se emplean con varones adultos (H), mujeres adultas (M), niños varones (V), niñas (N) o en alguna combinación (p. ej., VH). En consecuencia, exactamente 20 láminas son adecuadas para cada examinando.

En la figura 13–5 se muestra una ilustración similar a las utilizadas en la TAT. Al aplicar esta prueba, el examinador le pide a la persona examinada que invente una historia dramática para cada lámina, en la que explique aquello que condujo a la escena actual, qué sucede en ese momento, qué piensan y sienten los personajes y cuál será el desenlace. El examinador escribe la historia al pie de la letra para posterior calificación y análisis.

Henry Murray y sus colaboradores elaboraron el TAT en la Clínica psicológica de Harvard (Morgan y Murray, 1935; Murray, 1938). La prueba se diseñó originalmente para evaluar constructos como necesidades y presiones, que son elementos centrales de la teoría de personalidad de Murray. Según este autor, las necesidades organizan la percepción, pensamiento y acción y dan la energía a la conducta en dirección de su satisfacción. Ejemplos de necesidades incluyen aquéllas de logro, de afiliación y de dominio. En contraste, las presiones se refieren al poder de los acontecimientos ambientales para influir sobre la persona. La presión alfa la forman las fuerzas externas objetivas o "reales", mientras que la presión beta tiene que ver con los componentes subjetivos o percibidos de las fuerzas externas. Murray (1938, 1943) desarrolló un elaborado sistema de calificación del TAT para medir 36 necesidades diferentes y diversos aspectos de la presión, como se revelan en las historias de la persona examinada.

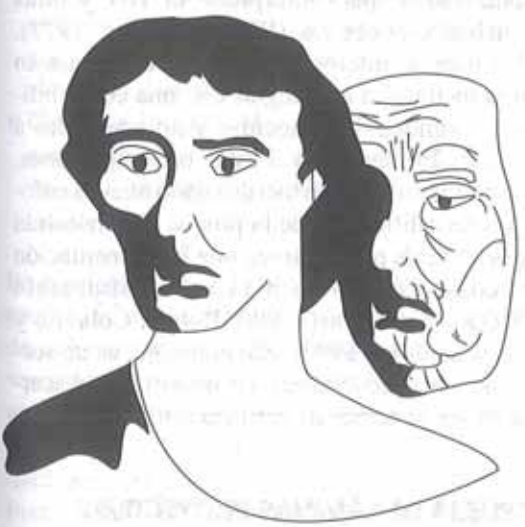


Figura 13–5. Ilustración similar a aquéllas del Test de apercepción temática.

Casi tan pronto como Murray publicó el TAT, otros clínicos comenzaron a desarrollar sistemas alternativos de calificación (p. ej., Dana, 1959; Eron, 1950; Shneidman, 1951; Tomkins, 1947). La literatura sobre aplicación, calificación e interpretación del TAT floreció de manera extensa, como lo documentan las revisiones recientes (Aiken, 1989, cap. 12; Groth-Marnat, 1997; Ryan, 1987; Weiner y Kuehnle, 1998). Para el decenio de 1950, no existía un modo preferido de aplicación, como tampoco un solo sistema favorito de calificación y ningún método predilecto de interpretación, una situación que todavía existe hoy en día. Incluso, los clínicos varían en el fraseo de las instrucciones y por lo común eligen un subconjunto individualizado de láminas del TAT para cada cliente. De hecho, la ausencia de procedimientos estandarizados es tal que en realidad debería considerarse al TAT como un método y no como una prueba.

Es necesario mencionar que las instrucciones de Murray incluían una declaración de que el TAT era "una prueba de la imaginación, la cual es una forma de inteligencia" y además estipulaba que:

> Voy a mostrarle algunas ilustraciones, una a la vez, y su tarea consistirá en inventar una historia tan dramática como pueda para cada una. Diga qué ha conducido al acontecimiento que se muestra en la imagen, describa qué está sucediendo en el momento, qué sienten y piensan los personajes y después diga el desenlace. Exprese sus pensamientos como le vengan a la mente. ¿Comprende usted? Ya que tiene 50 minutos

para las 10 ilustraciones, le puede dedicar cerca de cinco minutos a cada historia (Murray, 1943).

En la actualidad, los clínicos le han restado importancia al énfasis en la imaginación y la inteligencia cuando dan las instrucciones. Con toda seguridad esta omisión debe influir en la calidad de las historias producidas.

Aunque se han propuesto más de una docena de sistemas de calificación, en general la interpretación del TAT se basa en un análisis clínico-cualitativo de las producciones. Una consideración central tiene su origen en la suposición de Murray acerca del "héroe". Según esta perspectiva, éste es el protagonista de la historia de la persona examinada. Se supone que el individuo se identifica claramente con su personaje y proyecta en él sus propias necesidades, esfuerzos y sentimientos. A la inversa, los pensamientos, sentimientos o acciones que evita el héroe pueden representar áreas de conflicto para la persona examinada. Un ejemplo específico ayudará a aclarar estos puntos. Considere la respuesta que dio una persona deprimida ante la Lámina 3VH:*

> Parece como... no sé si es una chica o un chico. Podría ser cualquiera. Creo que no importa. Esta persona está muy agotada físicamente. Creo que es una ella. Está simplemente cansada. No ha quedado traumatizada ni nada. Sólo está sentada a la mesa con sus amigos y se siente realmente cansada. No está en ningún peligro de salud o nada. Éstas son sus llaves. Sus amigos la llevan arrastrando hasta su habitación y la meten en cama. Al día siguiente está bien. No queda traumatizada. Está cansada a nivel físico, no mental (Ryan, 1987).

Lo que destaca en la respuesta es la negación repetida del peligro o el Trauma. Pero después durante la prueba ya no se sostiene la negación de éste. Véase cómo respondió esta persona a la lámina en blanco, cuando relató una historia acerca de un joven, traumatizado en la escuela, que lleva su automóvil junto al río:

> Ve el puente, se siente realmente triste. Recuerda que ha escuchado historias sobre gente que salta y se mata. Nunca pudo comprender por qué lo hicieron. Ahora comprende, salta y se mata... debería haber esperado porque las cosas siempre se ponen bien a veces. Pero no esperó, se murió (Ryan, 1987).

La mayoría de los clínicos concluiría que la persona que produjo estas historias ha sido traumatizado y se defiende contra sus impulsos autodestructivos. De manera correspondiente, el clínico haría bien en explorar estas cuestiones en psicoterapia.

Es difícil valorar la idoneidad psicométrica del TAT debido a la abundancia de métodos de calificación e interpretación. Los clínicos defienden la prueba sobre bases anecdóticas, señalando las conclusiones notables y confirmatorias como la que se ilustra antes. No obstante, los investigadores que se inclinan hacia los datos se muestran más cautos. Un problema es que los protocolos del TAT que se califican de manera formal poseen una confiabilidad *test-retest* muy baja, con un valor mediano de $r = .28$ (Winter y Stewart, 1977). Lo que es más, 82% de los usuarios de la prueba emplean procedimientos "personalizados" para interpretar el TAT y otras pruebas proyectivas (Wade y Baker, 1977). Entonces, la interpretación del TAT se basa en gran medida en estrategias con una confiabilidad y validez desconocidas y no sometidas a prueba. En respuesta a estas preocupaciones, varios investigadores han diseñado nuevos enfoques de calificación de la prueba que muestran esperanza de proporcionar una fundamentación psicométrica sólida para este instrumento (McGrew y Teglasi, 1990; Ronan, Colavito y Hammondtree, 1993). Sin embargo, es de sorprender el poco avance en la investigación acerca de los sistemas de calificación del TAT.

PRUEBA DE LÁMINAS PROYECTIVAS

La *Picture Projective Test*, PPT (Prueba de láminas proyectivas), es un intento largamente

* La lámina 13VH representa a una persona —que puede ser varón o mujer— arrodillada o tirada a un lado de un sofá con la cabeza inclinada sobre un brazo. En una esquina se encuentra un objeto dibujado de manera vaga que algunos individuos examinados interpretan como un revólver o alguna otra arma.

esperado de construir un instrumento con propósitos generales que tuviera cualidades psicométricas mejoradas (Ritzler, Sharkey y Chudy, 1980; Sharkey y Ritzler, 1985). Los autores de esta prueba señalan que la mayor parte de las láminas del TAT ejercen un fuerte estímulo negativo como medios de "extracción" en la narración de historias. Dichas láminas se presentan en tonos oscuros, sombreados, y la mayor parte de las escenas representan personas en situaciones poco claras o sombrías. Por tanto, no es de sorprender que las respuestas proyectivas dadas a esta prueba se canalicen en gran medida hacia historias negativas y melancólicas (Goldfried y Zax, 1965).

En contraste, la PPT utiliza un nuevo conjunto de láminas tomadas del ensayo fotográfico llamado *Family of Man* que publicó el Museo de Arte Moderno (1955). Para la selección de las 30 fotografías se utilizaron los siguientes criterios:

- Las ilustraciones tenían que mostrar la posibilidad de producir material proyectivo significativo.
- La mayor parte, pero no todas las láminas, tenían que incluir más de un personaje humano.
- Cerca de la mitad de éstas tenían que representar a seres humanos que mostraran la expresión de afectos positivos (p. ej., sonreír, abrazarse, bailar).
- Cerca de la mitad de las fotografías tenían que representar seres humanos en poses activas, no simplemente parados, sentados o acostados.

En el piloteo inicial, los autores compararon las producciones de historias para el TAT y la PPT de ocho estudiantes de pregrado en diversas variables como extensión de las historias, tono emocional y nivel de actividad (Ritzler, Sharkey y Chudy, 1980). En comparación con las producciones para el TAT, las historias de la PPT fueron de extensión paralela, pero fueron mucho más positivas en cuanto a contenido temático y tono emocional. También fueron mucho más activas, lo cual significa que el personaje central tenía un efecto activo, autodeterminado, sobre la situación en la historia. Además, las historias de la PPT colocaban mayor énfasis en los temas interpersonales que en los intrapersonales. En otras palabras, las historias de la PPT daban mayor importancia a los aspectos "sanos", adaptativos, de la personalidad que las producciones para el TAT.

Los autores de la PPT también compararon su instrumento con el TAT en un estudio sobre validez diagnóstica (Sharkey y Ritzler, 1985). Se compararon las producciones de historias para la PPT y el TAT de 50 individuos: normales, depresivos ambulatorios, depresivos hospitalizados, psicóticos hospitalizados con buenos antecedentes premórbidos y psicóticos hospitalizados con antecedentes premórbidos negativos (10 personas en cada grupo). Aunque el TAT y la PPT tuvieron en esencia igual capacidad para discriminar entre individuos normales y depresivos, la PPT fue superior en diferenciar a personas psicóticas de aquéllas normales y depresivas. En la prueba, los individuos depresivos narraron historias con un tono emocional más sombrío y aquéllos con psicosis realizaron mayor cantidad de distorsiones perceptuales y desviaciones temáticas/interpretativas. La PPT parece ser un instrumento muy prometedor, aunque es obvio que se requiere más investigación sobre sus cualidades psicométricas. Un aspecto que es digno de mención es que cualquier persona puede comprar los estímulos de la PPT en una librería local. Los materiales requeridos se encuentran en la colección fotográfica *Family of Man* (Museo de Arte Moderno, 1955).

TEST DE APERCEPCIÓN INFANTIL

Diseñado como una extensión directa del TAT, el Test de Apercepción Infantil (CAT) consiste en 10 láminas y es adecuado para niños de 3 a 10 años de edad. La versión preferida para niños pequeños (CAT-A) representa animales en ambientes sociales indudablemente humanos (Bellak y Bellak, 1991). Los autores de la prueba utilizaron dibujos de animales bajo la supo-

sición de que los niños pequeños se identificarían mejor con éstos que con seres humanos. Se encuentra disponible una versión con figuras humanas (CAT-H) para niños de mayor edad (Bellak y Bellak, 1994). No existe un sistema formal de calificación para el CAT y no se proporciona información estadística sobre confiabilidad o validez. En vez de ello, el examinador prepara un diagnóstico o descripción de la personalidad con base en una síntesis de 10 variables registradas para cada historia: 1) tema principal; 2) héroe principal: 3) principales necesidades e impulsos del héroe; 4) concepto del ambiente (o mundo); 5) percepción de las figuras parentales, contemporáneos y personas más jóvenes; 6) conflictos; 7) ansiedades; 8) defensas; 9) idoneidad del superyó y 10) integración del yo (incluyendo originalidad de la historia y naturaleza del desenlace) (Bellak, 1992). La falta de atención a las cuestiones psicométricas en cuanto a calificación, confiabilidad y validez del CAT es perturbadora para la mayoría de los especialistas en pruebas.

OTRAS VARIACIONES SOBRE EL TAT

El TAT ha inspirado varias pruebas similares dirigidas a niños y adultos mayores (cuadro 13–7). Además, se han desarrollado modificaciones y variaciones de la prueba para minorías étnicas, raciales y lingüísticas. Una de las primeras fue el TAT de Thompson (T-TAT) en la que se volvieron a dibujar 21 láminas originales del TAT con figuras afroestadounidenses (Thompson, 1949). Esta modificación para el TAT incorporó ciertos cambios no intencionales —por ejemplo, en expresiones faciales y en las situaciones representadas. Como resultado, la T-TAT debería considerarse una prueba nueva y no sólo una traducción del TAT adecuada para personas afroestadounidenses (Aiken, 1989).

Otra prueba especializada similar al TAT es la TEMAS, que consiste de 23 láminas cromáticas que representan personas de origen hispano en interacción en ambientes urbanos contemporáneos (Aiken, 1989; Constantino, Malgady y Roger, 1988). TEMAS tiene el sig-

Cuadro 13–7. Pruebas de Apercepción Temática para Poblaciones Específicas

***Adolescent Apperception Cards* (Láminas de apercepción para adolescentes)**

Ésta es la única prueba de apercepción temática diseñada específicamente para adolescentes (12 a 19 años de edad). Las 11 láminas representan cuestiones contemporáneas relacionadas con los adolescentes; los temas incluyen soledad, estilos parentales, violencia intrafamiliar, actividad en bandas delictivas y abuso de drogas (Silverton, 1993). Los problemas con este instrumento incluyen los temas negativos que se representan en las láminas (que impiden asociaciones positivas) y la ausencia de cualquier tipo de enfoque objetivo para la calificación. Como ocurre con muchas técnicas de apercepción temática, la AAC es en realidad una herramienta clínica ideográfica, no una prueba.

Láminas Blacky

Para niños de cinco años de edad y mayores, la prueba de las Láminas Blacky también se basa en la premisa de que los niños se identifican con mayor facilidad con los animales que con seres humanos. Los 11 estímulos de caricatura representan las aventuras de un perro llamado Blacky y su familia (mamá, papá y su hermano Tippy). Además de requerirse una historia para cada lámina, el examinador también presenta una serie de preguntas de opción múltiple basadas en las etapas del desarrollo psicosexual derivadas de la teoría psicoanalítica (Blum, 1950). Aunque la prueba se desarrolló en un principio con adultos, los niños disfrutan de tomar a Blacky y responden bastante bien a las láminas. Algunos problemas con esta prueba incluyen la ausencia de normas, en especial para niños, y la baja estabilidad de las puntuaciones (LaVoie, 1987).

Prueba Pictórica Michigan–Revisada

Para niños entre 8 y 14 años de edad, la MPT-R consiste de 15 ilustraciones y una lámina en blanco. Las respuestas se califican en cuanto al Índice de tensión (p. ej., representación de idoneidad personal), Dirección de la fuerza (el que la figura central actúe o reciba los efectos de una acción sobre ella) y Tiempo verbal (p. ej., pasado, presen-

Cuadro 13–7. Pruebas de Apercepción Temática para Poblaciones Específicas (continuación)

te, futuro). Estas tres calificaciones se pueden combinar para producir un Índice de desadaptación. La confiabilidad y las normas son adecuadas, aunque la evidencia de su validez es insatisfactoria. Uno de los principales problemas con esta prueba es que las láminas presentan las relaciones interpersonales de manera tan vívida que literalmente se le deja poco a la imaginación del niño (Aiken, 1989).

Prueba de Apercepción para Edades Avanzadas (SAT)

Aunque las 16 situaciones representadas en las láminas de la SAT incluyen algunas circunstancias positivas, la mayoría de las ilustraciones se diseñaron para reflejar temas de desamparo, abandono, discapacidad, problemas familiares, soledad, dependencia y baja autoestima (Bellak, 1992). Los críticos se quejan de que la SAT estereotipa a los ancianos y, por tanto, desalienta las respuestas activas (Schaie, 1978; Klopfer y Taulbee, 1976).

nificado de "tema" en español y al mismo tiempo representa las siglas en inglés para "*tell me a story*" (cuéntame una historia). El contenido temático de las historias para la TEMAS se califica en cuanto a 18 funciones cognitivas, nueve de personalidad (yoicas) y siete afectivas. La prueba también puede calificarse en relación con varios índices objetivos como tiempo de reacción, fluidez, preguntas no respondidas y transformaciones del estímulo (p. ej., una carta se transforma en una bomba). Los niños de origen hispano responden bien a la TEMAS, aunque puedan mostrarse incapaces de expresarse ante las pruebas proyectivas tradicionales.

La confiabilidad inconsistente de la TEMAS es una fuente de gran preocupación, dado que la confiabilidad limita la validez. El manual informa que los resultados del coeficiente alfa de Cronbach para las 34 funciones de calificación fueron de .31 a .98, con la mitad por debajo de .70. Las confiabilidades *test-retest* fueron incluso más bajas; la correlación más elevada fue de $r = .53$ y para 26 de las 34 funciones las correlaciones ¡fueron casi cercanas a cero! A pesar de la cuestionable confiabilidad del instrumento, varios estudios proporcionan sustentación para su validez concurrente y predictiva. Por ejemplo, en una muestra clínica de 210 niños puertorriqueños, las calificaciones escalares de la TEMAS predijeron de manera razonablemente adecuada criterios independientes de desarrollo del yo, ansiedad rasgo y conducta adaptativa, con correlaciones entre .27 y .51 (Malgady, Constantino y Rogler, 1984). Un flujo constante de investigación ha continuado reforzando la utilidad de este instrumento, como lo revela el examen de Constantino y Malgady (1996). Ritzler (1993) proporciona una reseña en apoyo para la prueba, mientras que otros autores recomiendan que la TEMAS se utilice sólo con propósitos de investigación.

TÉCNICAS EXPRESIVAS

PRUEBA DEL DIBUJO DE LA FIGURA HUMANA

Como recordará el lector, en un capítulo anterior se dijo que Goodenough (1926) utilizó la tarea del dibujo de la figura humana como base para estimar la inteligencia. Enseguida, los psicólogos con inclinaciones psicodinámicas adaptaron el procedimiento para la evaluación proyectiva de la personalidad. Karen Machover (1949, 1951) fue la pionera en este nuevo campo. Su procedimiento se llegó a conocer como la Prueba del Dibujo de la Figura Humana (DFH). Este procedimiento disfrutó de popularidad desde un principio y aún se le utiliza ampliamente como herramienta de evaluación clínica. Watkins, Campbell, Nieberding y Hallmark (1995) informaron que los dibujos proyectivos como el DFH alcanzan el octavo lugar de popularidad entre los clínicos en EUA.

El DFH se aplica al presentarle al examinando una hoja de papel en blanco y un lápiz con borrador; pidiéndole que "dibuje una persona". Cuando se completa el dibujo, en gene-

ral se pide al individuo que dibuje a otra persona del sexo opuesto a la del primer dibujo. Por último, se le pide que "invente una historia sobre la persona como si él (o ella) fuera un personaje de una novela o de una obra de teatro" (Machover, 1949).

La interpretación del DFH procede de manera completamente clínica-intuitiva, guiada por varias hipótesis tentativas con base psicodinámica (Machover, 1949, 1951). Por ejemplo, Machover sostenía que es probable que las personas examinadas proyecten sus impulsos aceptables sobre la figura del mismo sexo y aquéllos inaceptables sobre la del sexo opuesto. También creía que los tamaños relativos de las figuras del varón y de la mujer revelaban indicios de la identificación sexual del individuo evaluado. Varias de las hipótesis interpretativas de Machover aparecen en el cuadro 13–8.

Estas premisas interpretativas son pintorescas, interesantes y plausibles. No obstante, se basan por completo en la teoría psicodinámica y en observaciones anecdóticas; Machover hizo pocos esfuerzos por validar las interpretaciones. El apoyo empírico para sus hipótesis es entre escaso y nulo (Swensen, 1957, 1968). En favor del DFH debe decirse que la calidad general de los dibujos predice, aunque de manera débil, la adaptación psicológica (Lewinsohn, 1965; Yama, 1990). Sin embargo, a juzgar por las normas contemporáneas para la evidencia, las evaluaciones radicales y displicentes de la personalidad que se obtienen del DFH con tanta frecuencia son vergonzosas. Algunos críticos han concluido que el DFH es una prueba deficiente que ya no debería utilizarse (Gresham, 1993; Motta, Little y Tobin, 1993).

Más que emplear el DFH para inferir los matices de la personalidad, una aplicación más apropiada de esta prueba es la detección de niños de quienes se sospecha un trastorno conductual o emocional. Para tal propósito, Naglieri,

Cuadro 13–8. Interpretaciones ilustrativas de la Prueba del dibujo de la figura humana

Signo	Significado interpretativo hipotético
Cabeza grande de manera desproporcionada	Enfermedad cerebral orgánica; cirugía cerebral previa; preocupaciones con dolores de cabeza.
Omisión de rasgos faciales deliberada	Evasión acerca de relaciones interpersonales muy conflictivas.
Boca dibujada con una raya gruesa	Personalidad verbalmente agresiva, demasiado crítica y en ocasiones sádica.
Mandíbula cambiada, borrada o reforzada	Compensación por debilidad, indecisión y temor a la responsabilidad.
Ojos masculinos de gran tamaño con pestañas	Varón con inclinaciones homosexuales, con frecuencia muy extravertido.
Énfasis en el cabello; p. ej., una barba	Una indicación de esfuerzos por la virilidad.
Énfasis gráfico en el cuello	Perturbación por la falta de control sobre los impulsos.
Tratamiento conspicuo del dedo índice, pulgar	Preocupación por la masturbación.
Indicación anatómica de órganos internos	Se encuentra sólo en la esquizofrenia o en pacientes con manía activa.

Fuente: Basado en Machover, K. (1949). *Personality projection in the drawing of the human figure.* Springfield, IL: Charles C. Thomas.

McNeish y Bardos (1991) desarrollaron el Dibujo de la figura humana: procedimiento de detección para el trastorno emocional (DAP: SPED). En un estudio, la precisión diagnóstica de los niños con problemas se mejoró de manera significativa a través de la aplicación del enfoque de calificación del DAP:SPED (Naglieri y Pfeiffer, 1992).

TÉCNICA DEL DIBUJO PROYECTIVO DE CASA-ÁRBOL-PERSONA (H-T-P)

La H-T-P es una prueba proyectiva que utiliza dibujos a pulso de una casa, un árbol y una persona (Buck, 1948, 1981). A la persona examinada se le da casi completa libertad de trazar los tres objetos, se piden dibujos a lápiz y a colores por separado. Aunque el examinador puede improvisar una prueba H-T-P con meras hojas de papel en blanco, Buck (1981) recomienda el uso de una forma de dibujo con cuatro hojas que tiene en la primera página la información de identificación. Las páginas 2, 3 y 4 tienen como título Casa, Árbol y Persona. Se necesitan dos formatos de dibujo para cada examinando, uno para los dibujos a lápiz y el otro para los dibujos a color. Buck (1981) también proporciona un formato independiente de cuatro hojas para una fase de interrogatorio posterior a los dibujos que consiste de 60 preguntas diseñadas para obtener las opiniones de la persona examinada en cuanto a los elementos de los dibujos. Muchos profesionales consideran que la fase de interrogatorio no amerita un esfuerzo adicional. También se cuestiona el valor de los dibujos realizados con colores (Killian, 1987).

La Técnica Proyectiva del Dibujo de casa-árbol-persona tiene en mucho la misma herencia que la Prueba del Dibujo de la Figura Humana. Como la Prueba DFH, la H-T-P se concibió en un principio como una medida de inteligencia, completa con un sistema de calificación cuantitativa para estimar un nivel aproximado de capacidad (Buck, 1948). Sin embargo, los clínicos pronto abandonaron el empleo de la H-T-P como medida de inteligencia y ahora se le utiliza casi de manera exclusiva como medida proyectiva de la personalidad.

Aunque no se ahondará en los detalles, la interpretación de la H-T-P depende de tres suposiciones generales: el dibujo de la Casa refleja la vida hogareña y las relaciones intrafamiliares de la persona examinada, el del Árbol refleja la manera en que el individuo experimenta el ambiente y el de la Persona hace eco de sus relaciones interpersonales. Buck (1981) proporciona numerosas hipótesis interpretativas tanto para los aspectos cuantitativos como cualitativos de los tres dibujos.

La H-T-P es una atractiva prueba que ha fascinado a los clínicos por más de 40 años. Por desgracia, Buck (1948, 1981) nunca ha proporcionado algún tipo de evidencia que sustente la confiabilidad o validez de este instrumento. De hecho, quizás él mismo es su peor crítico. En cierto punto dentro de su manual de la prueba, incluso llega a afirmar que la investigación sobre validación no es posible con la H-T-P (Buck, 1981, p. 164). Entre los impedimentos para dicha investigación cita los siguientes elementos:

1. Ningún signo en sí mismo es indicación infalible de alguna fortaleza o debilidad en el individuo.
2. Ningún signo en la H-T-P tiene un solo significado.
3. La importancia de un signo puede diferir notablemente de una constelación a otra.
4. La cantidad de datos diagnósticos y pronósticos que pueden obtenerse de cada uno de los puntos de análisis puede variar en gran medida de una persona a otra.
5. Los colores no tienen algún significado absoluto y universal.
6. Ninguno de los elementos del sistema de calificación cuantitativa puede tomarse automáticamente por su valor aparente (Buck, 1981).

En general, los intentos por validar la H-T-P como medida de personalidad han fallado (para una revisión véase Ellis, 1970; Hayworth, 1970;

Krugman, 1970; Killian, 1987). Críticos escrupulosos han recomendado en forma repetida que se abandone el uso de la H-T-P y enfoques similares de dibujo de figuras para la evaluación de la personalidad. Pero estos pronunciamientos se han topado en apariencia con oídos sordos. La popularidad de la H-T-P y de otras técnicas proyectivas continúa sin disminuir. En la sección final de este capítulo se ofrecen algunas reflexiones acerca de la aceptación que todavía tienen las técnicas proyectivas.

REITERACIÓN: LA PARADOJA PROYECTIVA

La evidencia es bastante clara en cuanto a que las inferencias obtenidas de las pruebas proyectivas son, con frecuencia, incorrectas. Ante las conclusiones negativas de validación, la duradera aceptación de los profesionales hacia estas pruebas constituye lo que se ha llamado la paradoja proyectiva. ¿Cómo se explica que instrumentos cuya evidencia de validez es, en el mejor de los casos, confusa, con frecuencia marginal, en ocasiones no existente o, incluso, decisivamente negativa, continúen siendo populares?

Se ofrecen dos explicaciones para la paradoja proyectiva. La primera es que los seres humanos se adhieren a estereotipos preexistentes aunque se les exponga a resultados contradictorios. Hace decenios, Chapman y Chapman (1967) demostraron este fenómeno con las pruebas proyectivas, llamándolo **validación ilusoria**. Estos investigadores le pidieron a estudiantes universitarios que observaran varios dibujos de figuras humanas similares a los obtenidos en la Prueba del dibujo de la figura humana (DFH). Los estudiantes ignoraban todo acerca de las pruebas proyectivas y no conocían las hipótesis interpretativas tradicionales para la DFH. Cada dibujo se acompañaba de descripciones breves de dos síntomas que supuestamente caracterizaban al paciente que había realizado los dibujos. En realidad, éstos se asignaron de manera aleatoria a los dibujos y consistían de trozos de tradición clínica asociada con la DFH que se había obtenido de un cuestionario que se había enviado antes por correo a varios psicólogos clínicos. Por ejemplo, dos de los síntomas utilizados eran:

1. Está preocupado con su hombría.
2. Sospecha de otras personas.

Cada estudiante recibió una combinación diferente de dibujos y se les asignaron los síntomas de manera aleatoria.

Después se les pidió que demostraran aquello que habían descubierto al describir, para varios dibujos, los síntomas que habían observado que se vinculaban con ese tipo de dibujo. Por supuesto, en realidad no había nada que descubrir, dado que síntomas y dibujos se combinaron aleatoriamente. No obstante, los participantes respondieron en términos de los estereotipos clínicos populares (p. ej., ojos poco comunes indican suspicacia, cabeza grande sugiere preocupación por la inteligencia). En apariencia, los estereotipos de sentido común que tenían los participantes surgieron fuertes y sin ningún daño —a pesar de la abundancia de ejemplos que no los confirmaban. Quizás algo similar ocurre en todos los campos de las pruebas proyectivas: los clínicos observan los casos confirmatorios, pero ignoran las conclusiones más numerosas que contradicen sus expectativas.

La segunda explicación para la paradoja proyectiva es que muchos clínicos no utilizan de ninguna manera los métodos proyectivos como pruebas, sino como enfoques auxiliares para le entrevista clínica. Éstos emplean las técnicas proyectivas como herramientas clínicas para obtener hipótesis tentativas acerca de la persona evaluada. La mayor parte de estas hipótesis resultarán falsas cuando se les examine de manera más cercana. Sin embargo, las pocas que se confirman pueden tener implicaciones importantes para el manejo clínico del individuo evaluado. Lo que es más, se sospecha que estas hipótesis fructíferas podrían no surgir —o podrían hacerlo con mayor lentitud— si el profesional se basara por completo en la

EJEMPLO DE CASO 13–1
PRUEBAS PROYECTIVAS COMO AUXILIARES DE LA ENTREVISTA

Un ejemplo específico puede ayudar a aclarar el papel de las técnicas proyectivas como auxiliares de la entrevista clínica. Durante la guerra de Vietnam, un psicólogo de la *Veteran's Administration* examinó a un joven soldado que por accidente se había disparado en una pierna con una pistola calibre .45, cuando practicaba cómo desenfundar con rapidez mientras se encontraba en la selva. Los cirujanos encontraron necesario amputarle la pierna desde la rodilla. El soldado estaba bastante deprimido y todos supusieron que sufría del duelo y la culpa por su enorme tragedia personal. Hablaba muy poco y se mostraba casi inestable. No obstante, se le persuadió para que realizara una serie de dibujos de figuras. En uno de ellos se representó a sí mismo como artillero de un helicóptero, que disparaba balas de manera indiscriminada a la selva debajo de sí. Cuando se le cuestionó sobre este dibujo, se empezó a comportar de modo muy animado y confesó que disfrutaba del combate. Guiado por las posibles implicaciones de este mórbido dibujo, el psicólogo buscó enterarse más de las actitudes del veterano hacia el combate. En el curso de varias entrevistas, el veterano reveló que disfrutaba en particular de disparar a objetos en movimiento —animales, soldados, civiles— sin importarle a quién. De manera gradual se hizo claro que este joven veterano era un criminal de guerra en estado incipiente y que se encontraba deprimido debido a que su lesión le impediría regresar al frente. Sobra decir que esta información tuvo un fuerte impacto sobre el tono del informe psicológico.

entrevista o utilizara pruebas formales con confiabilidad y validez establecidas (ejemplo de caso 13–1). No obstante, es difícil examinar empíricamente esta afirmación. Se está abierto a la posibilidad de que las aplicaciones clínicamente exitosas de las técnicas proyectivas proporcionen mayor evidencia de la validación ilusoria.

RESUMEN

1. Las pruebas proyectivas se basan en la hipótesis proyectiva: las interpretaciones personales de estímulos ambiguos deben reflejar por obligación las necesidades, motivos y conflictos inconscientes de la persona examinada. Métodos proyectivos populares incluyen la Prueba de interpretación de formas de Rorschach, el Test de apercepción temática, las pruebas de frases incompletas y las pruebas de dibujo (p. ej., la Prueba del dibujo de la figura humana).
2. El Rorschach, publicado en 1921, consiste de 10 manchas de tinta casi simétricas. En cada una el examinador pregunta "¿Qué podría ser esto?" En la fase de encuesta, el examinador aclara cuáles aspectos de la mancha (p. ej., forma o color) representaron un papel en la creación de cada respuesta.
3. El método de calificación del Rorschach que prefiere Exner codifica cada respuesta en cuanto a localización, forma, movimiento humano, utilización del color, contenido y otras variables. Las calificaciones de resumen y las proporciones de las variables suministran las hipótesis sobre el funcionamiento de la personalidad. A pesar de que durante largo tiempo ha sido popular, todavía existen dudas alrededor de la prueba Rorschach con respecto a su confiabilidad y validez.
4. La *Holtzman Inkblot Test* incluye 45 láminas; se requiere una sola respuesta para cada una. Las categorías de calificación de la HIT

son sumamente confiables, con una concordancia intercalificadores que por lo general se localiza en el rango de .90. Los estudios de validez en los que se utilizan reglas simples de decisión sustentan el empleo de la HIT como un auxiliar para el psicodiagnóstico, en especial de la esquizofrenia.

5. El *Rotter Incomplete Sentences Blank* (RIBS) contiene 40 partes de frases escritas en su mayor parte en primera persona. Cada frase completada recibe una calificación de adaptación de 0 (buena) a 6 (deficiente); la suma forma la calificación total de ajuste. Las tasas de clasificación correcta (p. ej., adaptación contra desadaptación) son demasiado bajas para la toma de decisiones a nivel individual.
6. El *Rosenzweig Picture Frustration Study* (Estudio P-F) consiste de 24 dibujos, cada uno de los cuales muestra a dos personas en una situación sumamente frustrante. El individuo examinado proporciona la primera respuesta verbal que le viene a la mente. Las calificaciones objetivas indican los modos típicos de reacción ante la frustración. Debido a su poca confiabilidad, el Estudio P-F es adecuado primordialmente para la investigación.
7. El Test de apercepción temática (TAT) consiste de 30 dibujos y fotografías en blanco y negro; una de las láminas está en blanco. Se pide al examinando que invente una historia dramática para cada lámina, incluyendo el pasado, presente, futuro y los sentimientos de los personajes principales. En general, la interpretación del TAT depende del análisis clínico-cualitativo de las producciones de historias.
8. Variaciones del TAT incluyen la *Picture Projective Test*, la cual se basa en fotografías del ensayo fotográfico *Family of Man*; el TAT de Thompson para afroestadounidenses; el Test de apercepción infantil (CAT), el cual utiliza dibujos de animales; la TEMAS, una prueba de apercepción diseñada para personas de origen hispano; y la prueba de apercepción para personas de edad avanzada.
9. En la prueba de Dibujo de la figura humana (DFH) de Machover, simplemente se le pide a la persona examinada que "dibuje una persona". La interpretación se realiza de manera clínica-intuitiva con base en hipótesis publicadas; por ejemplo, remarcar el trazo de la mandíbula señala indecisión. Otro instrumento que sigue una línea similar es la Técnica del dibujo proyectivo de Casa-Árbol-Persona (la persona evaluada dibuja estos objetos), que tiene evidencia de validez también muy limitada.
10. La paradoja proyectiva (la popularidad duradera de las pruebas proyectivas a pesar de su cuestionable validez) puede explicarse, en parte, por el fenómeno de la validación ilusoria. Ésta se demuestra cuando los individuos ignoran los casos que no confirman sus conclusiones y se adhieren a sus estereotipos preexistentes.

TÉRMINOS Y CONCEPTOS CLAVE

Prueba proyectiva p. 593
Hipótesis proyectiva p. 594
Frustración p. 603
Validación ilusoria p. 612

CAPÍTULO 14

Evaluación estructurada de la personalidad

TEMA 14A
Inventarios de autoinforme

Inventarios guiados por la teoría
Inventarios derivados del análisis factorial
Inventarios con codificación de criterio
Resumen
Términos y conceptos clave

La historia de la evaluación de la personalidad puede caracterizarse por dos tendencias superpuestas. Primero, las técnicas proyectivas desestructuradas, como la prueba Rorschach, dominaron las pruebas de personalidad a inicios del siglo XX y después perdieron su popularidad. En segundo lugar, los enfoques estructurados como los inventarios de autoinforme y las clasificaciones conductuales obtuvieron importancia a mediados del siglo y luego aumentaron rápidamente su popularidad. En el tema anterior, se introdujo al lector a las muchas variedades de técnicas proyectivas. Estos métodos son resplandecientes en cuanto a las hipótesis que dan por resultado; sin embargo, las técnicas proyectivas carecen en gran medida de la aprobación de los clínicos con orientación psicométrica. Este capítulo se enfocará en los métodos más objetivos para la evaluación de la persona que favorecen los psicólogos con una mentalidad dirigida a la medición. En el tema 14A, Inventarios de autoinforme, se revi-

san los instrumentos con respuestas de cierto/falso y de elección forzosa, incluyendo la prueba de personalidad de más amplio uso, el Inventario multifásico de la personalidad (MMPI) y su revisión más reciente, el MMPI-2. En el tema 14B, Evaluación conductual y enfoques relacionados, se analizarán los enfoques más recientes que dependen de las observaciones y clasificaciones conductuales.

Los psicómetras contemporáneos han dependido de tres tácticas para el desarrollo de las pruebas: los enfoques ligados a la teoría, las estrategias analítico-factoriales y los métodos con codificación de criterio. Se organizará la discusión acerca de los inventarios de autoinforme alrededor de estas tres categorías. Por supuesto, las fronteras son un tanto artificiales y muchos de los autores de pruebas utilizan una combinación de métodos.

Los enfoques estructurados para el examen de la personalidad que se analizan en las siguientes secciones están empapados de los detalles de la metodología psicométrica. Estas pruebas presentan referencias notables a los índices de confiabilidad, codificación de criterio, análisis factorial, validación de constructo y otras formas de conocimiento técnico. Por esta razón, es frecuente que los enfoques que se analizan aquí se consideren objetivos —en contraste con los proyectivos. No obstante, el que se consideren objetivos en cualquier sentido significativo es en realidad una pregunta empírica que debe responderse con base en la investigación. Quizá sea más adecuado llamarlos métodos **estructurados**. Lo son en el sentido de que para su aplicación, calificación e interpretación se siguen reglas sumamente específicas. De hecho, algunos de los enfoques son tan absolutamente estructurados que la persona examinada puede responder a las preguntas presentadas en una pantalla de computadora y observar un informe narrativo generado por ésta, el cual se emite por la impresora unos segundos después.*

* Los informes narrativos por computadora pueden no ser del todo un desarrollo positivo. En el siguiente capítulo se analizan los beneficios y riesgos de los informes generados por computadora.

INVENTARIOS GUIADOS POR LA TEORÍA

La construcción de varios inventarios de autoinforme se ha guiado de manera estrecha por las teorías formales o informales de la personalidad. En estos casos, el autor de la prueba diseñó el instrumento alrededor de una teoría preexistente. Los inventarios guiados por la teoría contrastan con los enfoques analítico-factoriales, que con frecuencia producen una teoría retrospectiva basada en las conclusiones iniciales de la prueba. Estos inventarios también difieren del estricto empirismo ateórico que se encuentra en los instrumentos con codificación de criterio como el MMPI y el MMPI-2. Ejemplos de inventarios guiados por la teoría incluyen el *Edwards Personal Preference Schedule* (Inventario Edwards de preferencias personales; EPPS) y la *Personality Research Form* (Forma de investigación de la personalidad; PRF), ambos basados en la teoría de personalidad de Murray (1938) sobre necesidades-presiones. Otros ejemplos incluyen el Inventario Tipológico de Myers-Briggs (MBTI), que representa una aplicación de la teoría de Carl Jung sobre tipos de personalidad. El *Jenkins Activity Survey* (Examen Jenkins de actividad), diseñado para evaluar el patrón conductual de Tipo A con propensión a la cardiopatía, también tipifica un instrumento guiado por la teoría. Por último, algunos inventarios guiados por la teoría, como el Inventario de ansiedad: estado-rasgo (IDARE), intentan medir componentes muy específicos de la personalidad. A continuación se revisa en mayor detalle cada una de estas pruebas.

INVENTARIO EDWARDS DE PREFERENCIA PERSONAL

El *Edwards Personal Preferences Schedule*, EPPS (Inventario Edwards de Preferencia Personal), constituyó el primer intento de medir la teoría de Murray (1938) sobre necesidades manifiestas con un inventario estructurado de la personalidad (Edwards, 1959; Helms, 1983).

El lector recordará de un análisis anterior que Murray postuló 15 necesidades y desarrolló una prueba proyectiva, el Test de apercepción temática, para detectar dichas necesidades. Edwards, un psicómetra consumado muy conocedor de los matices de la teoría de la medición, buscó desarrollar una prueba objetiva, estructurada, para medir esas 15 necesidades de manera más confiable y válida. En la figura 14–1 se listan las 15 necesidades junto con un perfil del EPPS.

El EPPS consiste de 210 pares de afirmaciones en las que los reactivos de cada una de las 15 escalas se equiparan con los de las 14 restantes. El inventario emplea un formato de elección forzosa en el que la persona examinada debe elegir una afirmación de cada par que sea la más representativa a nivel personal. El formato de elección forzosa del EPPS es peculiar y resulta incómodo para la mayoría de las personas sometidas a prueba, debido a que con frecuencia los coloca, como diría el proverbio,

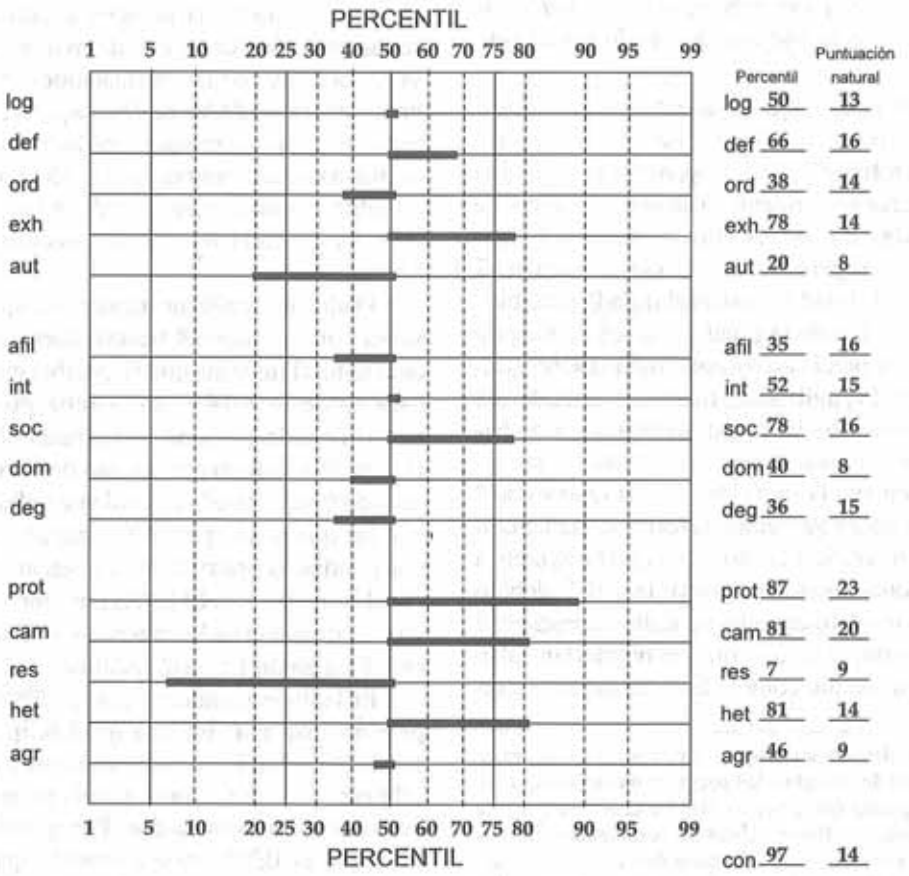


Figura 14–1. Perfil de muestra del Inventario Edwards de preferencias personales.
Fuente: Reproducido con autorización de Edwards, A. L. (1959). *Edwards Personal Preference Schedule.* Nueva York: Psychological Corporation. Copyright © 1954, 1959 de la Psychological Corporation. Reproducido con autorización. Derechos reservados.

entre la espada y la pared. Aquí se presentan tres reactivos similares a los del EPPS; para cada uno, la persona debe elegir aquella afirmación que sea más característica a nivel personal:

1. A. Me gusta hablar frente a un grupo.
 B. Me gusta trabajar para lograr las metas elegidas por mí mismo.
2. A. Me siento triste cuando veo una noticia trágica en la televisión.
 B. Me siento nervioso cuando tengo que hablar frente a un grupo.
3. A. No me importaría limpiar 10 galones de jarabe.
 B. No me importaría escalar un acantilado empinado con una cuerda de seguridad.

¿Por qué adoptó Edwards este incómodo formato en su prueba? La respuesta tiene que ver con el problema de la propensión a responder de maneras socialmente deseables. La **serie de respuestas de deseabilidad social** es la tendencia de las personas examinadas a responder a la deseabilidad (o indeseabilidad) percibida en un reactivo de la prueba, en vez de hacerlo de manera precisa a su contenido. Dicho simplemente, los individuos tienden a refrendar las afirmaciones socialmente deseables y a no hacerlo con las socialmente indeseables —sin tomar en cuenta la veracidad de las respuestas.* La mayoría de las personas responderían cierto a una afirmación como: "Me gusta ayudar a las personas ancianas a cruzar la calle", debido a que el reactivo aprueba un atributo deseable a nivel social, mientras que responderían falso a una afirmación como: "En ocasiones he tenido fantasías sobre la muerte de mis padres", debido a que el reactivo permite una cualidad socialmente indeseable. Pero en el caso de algunas personas, la respuesta socialmente deseable no es precisa en realidad. Después de todo, muchas personas no disfrutan en verdad de ayudar a los demás y la mayoría de los individuos han tenido fantasías sobre posibilidades desagradables.

La elegancia del EPPS es que los pares de afirmaciones en cada reactivo se equiparan en cuanto a deseabilidad social (Edwards, 1957). Debido a que cada afirmación en un par dentro del reactivo tiene el mismo nivel de deseabilidad social, el contenido de cada afirmación ejercerá mayor poder de "extracción" para determinar la elección de la persona examinada. Por supuesto, Edwards diseñó de manera cuidadosa el contenido de las afirmaciones para incorporar las necesidades de Murray. Al hacer que todas las comparaciones posibles entre pares de afirmaciones englobaran las 15 necesidades, el EPPS produce una medida de la fortaleza relativa de cada una de las necesidades de Murray.

Dado que cada necesidad se aparea dos veces con las otras 14 necesidades, la puntuación natural máxima que es posible en cada escala es de 2×14 o 28. Dicha puntuación ocurriría si la persona examinada eligiera la afirmación para una necesidad dada como más característica a nivel personal que todas las otras con las que se le aparea. Por supuesto, la menor puntuación posible en una escala de necesidad sería de cero. El EPPS también incorpora una verificación de la consistencia al repetir 15 reactivos en un formato idéntico.

El lector reconocerá que el EPPS es una **prueba ipsativa**. En una medida ipsativa, la calificación total —promediada entre todas las subpruebas— es siempre la misma para todos los individuos examinados. Por ejemplo, debido a que el EPPS es una medida ipsativa, la puntuación escalar promedio total es siempre 14, y las puntuaciones escalares altas deben contrabalancearse con las bajas. También debe recordarse que, en una escala ipsativa, las puntuaciones altas son relativas, no absolutas. En

* Aquí se debería mencionar que la predisposición de respuesta de deseabilidad social es una tendencia natural humana que se encuentra en casi toda persona examinada. La forma extrema de esta tendencia es el intento consciente y deliberado de "fingirse bueno" en las pruebas de personalidad. Pero la deseabilidad social no necesariamente indica un engaño deliberado de parte del individuo. Ésta siempre se halla presente en cierto grado, incluso cuando la persona evaluada adopta una perspectiva de "buena fe" al realizar la prueba de personalidad.

otras palabras, la fortaleza de cada necesidad no se expresa de manera absoluta sino relativa a la de otra necesidad de la persona examinada. Por tanto, produce confusión el hecho de que Edwards (1959) también recomendara presentar las calificaciones del EPPS en un formato normativo. El manual proporciona puntuaciones *T* y percentiles mediante los cuales pueden compararse las puntuaciones naturales de un individuo con los resultados de muestras universitarias y de adultos en general (véase figura 14–1). Al combinar dos metodologías incompatibles (ipsativa y normativa) en el mismo instrumento, Edwards hace que la interpretación de su prueba sea confusa.

El EPPS se utiliza ampliamente en la orientación universitaria como medio de descubrimiento personal y recibe un empleo ocasional como herramienta de investigación. No obstante, algunos críticos consideran al EPPS como un ejercicio en construcción de pruebas más que como un artículo serio en el mercado de las pruebas validadas (Heilbrun, 1972; Drummond, 1987). Muchos clientes se frustran y aburren cuando realizan la prueba. Lo que es más, la estandarización es anticuada y los datos de confiabilidad no son particularmente emocionantes. Por ejemplo, la confiabilidad *test-retest* de las 15 puntuaciones escalares va de .55 a .87, con una mediana de .73. Cooper (1990) concluyó que las normas presentadas en el trabajo de Edwards (1959) no corresponden a estudios normativos más recientes.

Los primeros intentos de validar el EPPS a través de calificaciones comparativas de la fortaleza de las necesidades de Murray con puntuaciones en el EPPS tuvieron resultados mixtos (Drummond, 1987). Sin embargo, un estudio reciente de Piedmont, McCrae y Costa (1992) proporciona fuerte apoyo a la validez del EPPS. Estos investigadores correlacionaron las calificaciones en el EPPS con aquéllas en el Inventario NEO de personalidad (NEO PI) para 330 personas en estudios de pregrado. El NEO PI, que se analiza después, mide cinco constructos: Neuroticismo, Extraversión, Apertura a la experiencia, Conformidad y Escrupulosidad. El patrón de relaciones sustentó la validez convergente de ambos instrumentos, con escalas que muestran correlaciones apropiadas y confirmatorias de la teoría. Por ejemplo, Agresión en el EPPS se correlacionó en .47 con Neuroticismo del NEO PI y en –.53 con Conformidad del NEO PI. Las relaciones fueron más fuertes y con mayor confirmación de la teoría cuando el EPPS se calificó de manera normativa. El formato ipsativo, de elección forzosa, del EPPS aparentemente disminuyó los coeficientes de validez y la validez convergente y de discriminación.

FORMA DE INVESTIGACIÓN DE LA PERSONALIDAD

Otra prueba basada en el sistema de necesidades de Murray es la *Personality Research Form*, PRF (Forma de Investigación de la Personalidad; Jackson, 1970, 1984b). Ésta se encuentra disponible en varias formas que difieren en el número de escalas o de reactivos por escala. Además de las pruebas cortas paralelas (formas A y B), la PRF también puede obtenerse como formas largas paralelas (formas AA y BB). Éstas, las cuales se utilizan principalmente con estudiantes universitarios, consisten de 440 reactivos de cierto/falso. Las formas largas producen 20 calificaciones escalares de personalidad y dos de validez, Infrecuencia y Deseabilidad (cuadro 14–1). La versión más popular de la PRF es la forma E, que consiste de las 22 escalas en una prueba modificada con 352 reactivos.

Para construir las formas largas de la PRF, Jackson desarrolló primero descripciones precisas y detalladas de los constructos a medir. A continuación, se escribieron más de 100 reactivos para cada escala a fin de detectar los rasgos subyacentes a las necesidades hipotéticas. Después de la revisión editorial, estos reactivos se aplicaron a grandes muestras de estudiantes universitarios. Con base en correlaciones biseriales elevadas con las calificaciones totales de escala y la poca correlación con las calificaciones de otras escalas y de la de Deseabilidad, se seleccionaron 20 reactivos para cada una.

Cuadro 14–1. Escalas de la Forma de Investigación de la Personalidad

Escala	Interpretación de la calificación elevada
Degradación	Modesto, humilde, acepta la culpa
Logro	Lucha por obtener las metas, competitivo
Afiliación	Amistoso, aceptante, sociable
Agresión	Discute, es combativo, se molesta con facilidad
Autonomía	Independencia, evita las restricciones
Cambio	Evita la rutina, busca el cambio
Estructura cognitiva	Prefiere la certidumbre, le desagrada la ambigüedad
Defensa	Está en guardia, se ofende con facilidad
Dominio	Influyente, disfruta del liderazgo
Resistencia	Perseverante, trabajador esforzado
Exhibición	Dramático, disfruta la atención
Evitación del daño	Evita riesgos y emociones
Impulsividad	Impulsivo, se expresa con libertad
Protección	Solícito, comprensivo, brinda consuelo
Orden	Organizado, le desagrada la confusión
Juego	Juguetón, despreocupado, disfruta de las bromas
Sensibilidad	Se percata y recuerda las sensaciones
Reconocimiento social	Se preocupa por la reputación y aprobación
Socorro	Inseguro, busca que lo tranquilicen
Comprensión	Valora el pensamiento lógico
Deseabilidad	Escala de validez: presentación favorable
Infrecuencia	Escala de validez: respuestas poco frecuentes

A diferencia de muchos otros inventarios de personalidad, las escalas PRF no tienen superposición de reactivos. Como resultado, las escalas tienen una independencia excepcional, con la mayor parte de los coeficientes de intercorrelación en las cercanías de ± .30 (Gynther y Gynther, 1976). Además, los rígidos procedimientos de construcción de escala que empleó Jackson (1970) produjeron escalas con una consistencia interna excepcionalmente buena, con un rango de .80 a .94 y mediana de .92. Un aspecto deseable de la PRF es su legibilidad: la prueba requiere sólo un nivel de lectura de quinto o sexto grados (Reddon y Jackson, 1989). MacLennan (1992) ha preparado una bibliografía comentada para la PRF con más de 375 citas.

La validez de constructo de la PRF depende en especial de análisis factoriales confirmatorios que corroboran la agrupación de los reactivos en 20 escalas (Jackson, 1970, 1984b). Además, la investigación indica correlaciones positivas con escalas comparables sobre otros inventarios (Mungas, Trontel y Weingardner, 1981). Por ejemplo, Edwards y Abbott (1973) encontraron correlaciones excepcionalmente fuertes y confirmatorias entre escalas similares en la PRF y el *Edwards Personality Inventory* (Inventario de personalidad Edwards; EPI);

(Edwards, 1967). El EPI es una prueba respetada pero que se utiliza pocas veces y que consiste de 1 200 (!) preguntas de cierto/falso. A continuación se presentan algunas de las correlaciones confirmatorias entre las escalas de la PRF y del EPI para 218 estudiantes universitarios de ambos sexos:

Rendimiento (PRF) × Es un trabajador esforzado (EPI). .74

Cambio (PRF) × Le agrada una rutina establecida (EPI). –.54

Protección (PRF) × Ayuda a otros (EPI). .64

Socorro (PRF) × Dependiente (EPI). .73

Debido a que estos instrumentos se desarrollaron de manera independiente, según diferentes filosofías de construcción de pruebas, los datos refuerzan la validez de ambas pruebas.

INVENTARIO TIPOLÓGICO DE MYERS-BRIGGS (MBTI)

Publicada originalmente en 1962, el MBTI es un inventario de autoinforme, de elección forzosa, que intenta clasificar a las personas según una adaptación de la teoría de tipos de personalidad de Carl Jung (Myers y McCaulley, 1985; Tzeng, Ware y Chen, 1989). El instrumento se presenta en una versión con 166 reactivos (Forma F) y una versión con 126 reactivos (Forma G). Aquí se analiza principalmente la Forma F debido a que es la que más se utiliza. De hecho, es posible que el MBTI sea la prueba de personalidad que más se utiliza con poblaciones psiquiátricas (DeVito, 1985).

El MBTI se califica en cuatro dimensiones teóricamente independientes: Extraversión-Introversión, Sensación-Intuición, Pensamiento-Sentimiento, Juicio-Percepción. Aunque las calificaciones en cada dimensión bipolar son continuas, es práctica común que las de una persona examinada se resuman de manera tipológica. Por ejemplo, un individuo podría obtener calificaciones más inclinadas hacia la Extraversión, Intuición, Sentimiento y Percepción y, por tanto, obtener un tipo resumido de ENSP.* Dicho perfil sugeriría las siguientes características de personalidad; una mayor relación con el mundo exterior de las personas y cosas que con el mundo interior de las ideas (E); una tendencia a buscar las posibilidades en vez de trabajar con hechos conocidos (N); una inclinación a basar los juicios en valores personales más que en el análisis y la lógica (S) y una preferencia por un modo de vida flexible y espontáneo más que por una existencia ordenada y planeada (P).

Los datos de estandarización del MBTI consisten de normas percentiles para las calificaciones de los cuatro indicadores, derivadas de pequeñas muestras de estudiantes de preparatoria y universidad. Las confiabilidades por mitades para las cuatro escalas se encuentran en los rangos de .70 y .80. Quizás en parte debido a que los estudios en apoyo de su validez son escasos, la interpretación tipológica del MBTI no ha sido bien recibida en términos generales por los psicólogos orientados a la medición. Un problema con este enfoque es que la interpretación parece demasiado superficial y sencilla y posee una calidad tipo horóscopo. En justicia con el MBTI, existen maneras más elaboradas de interpretar el instrumento, como lo revela la explosión de investigación reciente. Kaufman, McClean y Lincoln (1996) listan referencias relacionadas.

MEDIDAS DE LA CONDUCTA TIPO A

Se encuentran disponibles varias medidas tipo cuestionario de la conducta Tipo A, creadas con propósitos de investigación. La más reciente es la *Time Urgency and Perpetual Activation*; TUPA (Escala de Premura y Actividad Continua; Wright, McCurdy y Rogoll, 1992). La escala TUPA es respetada por investigadores en medicina conductual debido a su excelencia psicométrica. No obstante, el instrumento más conocido y más utilizado de este tipo es el Examen Jenkins de actividad.

* Dado a que existen dos polos para cada una de las cuatro dimensiones, el número de posibles tipos de personalidad es 2^4, que es 16.

La *Jenkins Activity Survey*, JAS (Encuesta Jenkins de Actividad), es un cuestionario de autoinforme con 52 reactivos de opción múltiple diseñados para identificar el patrón conductual Tipo A, con propensión a la cardiopatía, que se analizó en el capítulo anterior (Jenkins, Zyzanski y Rosenman, 1979). Los reactivos en la JAS se parecen a los siguientes:

En la actualidad, usted considera que es:

A. ¿Definitivamente competitivo y ambicioso?
B. ¿Probablemente competitivo y ambicioso?
C. ¿Probablemente más relajado y despreocupado?
D. ¿Definitivamente más relajado y despreocupado?

Además de la puntuación mixta de la conducta Tipo A, la JAS produce tres subescalas derivadas de manera analítico-factorial: Velocidad e impaciencia, Participación laboral y Dinamismo/Competitividad. Las correlaciones entre la puntuación mixta de la escala Tipo A y las tres subescalas son modestas (.42 a .67), lo cual indica que las puntuaciones factoriales pueden proporcionar contribuciones independientes a la evaluación de las tendencias Tipo A. Las normas de la JAS se obtuvieron con 2 588 empleados varones de clase media entre 48 y 65 años de edad. El instrumento se estandarizó para tener una media de 0 y una desviación estándar de 10.0, donde las calificaciones positivas indican tendencias Tipo A y las negativas indican lo contrario, tendencias Tipo B.

Con el propósito de una revisión breve se dirá que el patrón conductual Tipo A consiste de inseguridad de estatus, hiperagresividad, hostilidad libre flotante y un sentido de apremio de tiempo (Friedman y Ulmer, 1984). Algunos estudios indican que las personas con este patrón conductual tienen mayor riesgo de cardiopatía coronaria (CC). Por tanto, la identificación temprana de individuos en alto riesgo podría tener implicaciones portentosas para la intervención. Antes de la JAS, una entrevista estructurada prolongada proporcionaba el único medio para la identificación de las personas con el patrón conductual Tipo A. La JAS se desarrolló como un intento por duplicar la entrevista estructurada, con lo cual proporciona un método rápido y económico para detectar la conducta Tipo A.

Por desgracia, la JAS no ha satisfecho sus ambiciosas aspiraciones. La confiabilidad *test-retest* de las tres subescalas es marginal, cuando mucho, con valores tan bajos como .58 para Velocidad e impaciencia, .66 para Participación laboral y .71 para Dinamismo/Competitividad (Bishop, Hailey y O'Rourke, 1989; Igbokwe, 1989). Lo que es más, el nivel de concordancia entre la entrevista estructurada y las calificaciones de la JAS es sólo regular, sin alcanzar la magnitud suficiente como para justificar el uso de esta prueba para diagnóstico individual (Yarnold y Bryant, 1988). Otro problema con la JAS es que los pacientes con CC no difieren de los pacientes médicos en general en cuanto a las subescalas de la prueba. En una comparación entre 40 pacientes con CC y 40 con otros problemas médicos, Wright (1992) encontró que la escala de Velocidad e Impaciencia produjo una diferencia significativa y apropiada, pero Dinamismo/Competitividad produjo una diferencia significativa en dirección contraria —los pacientes con CC obtuvieron calificaciones más bajas que aquéllos sin CC.

Además, es obvio que las normas no son representativas, dado que no incluyen a mujeres, jóvenes o ancianos o a personas de niveles socioeconómicos bajos. La JAS también tiene dificultades en cuanto a su calificación manual, debido al complejo sistema de ponderación que se utiliza. Blumenthal (1985) ofrece una revisión poco halagadora de la JAS y recomienda que se le utilice sólo para la investigación clínica y experimental.

Sin embargo, los investigadores continúan encontrando utilidad en la JAS, aunque ahora se reconoce que las subcalificaciones específicas predicen mejor los problemas de salud que la calificación global (Hart, 1997).

INVENTARIO DE ANSIEDAD: RASGO-ESTADO (IDARE)

El IDARE es una medida corta y rápida de la ansiedad rasgo y estado que ha recibido altas

calificaciones debido a su mérito técnico (Spielberger y colaboradores, 1983). La **ansiedad estado** son los sentimientos transitorios de temor o preocupación que la mayoría experimenta en ciertas ocasiones. La **ansiedad rasgo** es la tendencia relativamente estable de un individuo a responder de manera ansiosa a una situación estresante. Estos dos constructos son independientes pero se relacionan: el nivel de ansiedad rasgo refleja la propensión a presentar ansiedad estado.

El IDARE es una medida con 40 reactivos que evalúa por separado ambos tipos de ansiedad. Los reactivos en la prueba son términos simples y descriptivos como "nervioso", "seguro" y "relajado". Los 20 reactivos de ansiedad estado se califican en una escala de intensidad cuatro puntos, clasificados de la siguiente manera: "De ninguna manera", "En cierto modo", "Moderadamente" y "Mucho". A las personas examinadas se les pide que califiquen estos reactivos según cómo se sientan "en este momento". Los 20 reactivos de ansiedad rasgo se califican en una escala de intensidad de cuatro puntos que se clasifican como "Casi nunca", "A veces", "Con frecuencia" y "Casi siempre". Se pide a las personas que califiquen estos reactivos según como "se sienten en general".

El IDARE se emplea con estudiantes de preparatoria, universitarios y adultos. Una versión similar para niños, el *State-Trait Anxiety Inventory for Children* (Inventario de ansiedad estado-rasgo para niños), se dirige a estudiantes de primaria y secundaria. Los aspectos técnicos de esta prueba son buenos en general y las muestras de estandarización son grandes y representativas. Por ejemplo, la confiabilidad *test-retest* se encuentra en la parte alta del rango de .70 para ansiedad rasgo y, como podría predecirse, es más baja para la ansiedad estado (.27 a .62 en diversos estudios). Deben mencionarse dos puntos débiles del IDARE. Primero, el constructo de la ansiedad rasgo no está bien definido y parece incluir rasgos no relacionados, como un sentimiento general de insatisfacción con uno mismo (Chaplin, 1984). Además, el IDARE es totalmente un instrumento con validez de Facie que puede falsearse con impunidad. Por tal razón, los resultados en el IDARE deben interpretarse con considerable cuidado, en particular cuando la situación demanda resultados "sanos" en la prueba. Spielberger (1984) ha publicado una bibliografía amplia de estudios de investigación con esta prueba.

INVENTARIOS DERIVADOS DEL ANÁLISIS FACTORIAL

CUESTIONARIO DE 16 FACTORES DE LA PERSONALIDAD (16FP)

El 16FP es una prueba de personalidad de elección forzosa que se utiliza ampliamente y que en la actualidad se encuentra disponible en cinco formas independientes. Cada una consiste de oraciones parciales de tipo declarativo que requieren que la persona examinada responda a una situación específica al elegir entre dos (Forma E) o tres opciones de elección forzosa (formas A, B, C y D). Ejemplos de reactivos del tipo de los del 16FP incluyen:

Tomo decisiones con base en
a. Sentimientos.
b. Sentimientos y razón por igual.
c. La razón.

¿Cuál de los siguientes elementos es diferente de los otros?*
a. Vela.
b. Estrella.
c. Lámpara incandescente.

Me resulta difícil dar un discurso ante personas desconocidas.
a. Sí.
b. En cierto modo.
c. No.

* Al lector puede parecerle curiosa la inclusión en el 16FP de lo que parecen ser reactivos de prueba de inteligencia. De hecho, desde hace largo tiempo los psicólogos han reconocido que la personalidad y la inteligencia se encuentran entrelazadas de manera compleja. La mayoría de autores de pruebas han atendido a este dilema haciendo el intento por separar personalidad e inteligencia. En vez de ello, Cattell decidió explorar la superposición entre personalidad e inteligencia al incluir elementos de ambas dentro de la misma prueba.

Las formas contienen de 105 a 187 reactivos y difieren principalmente en el nivel de lectura (de un nivel de tercer grado a uno de primero de secundaria). La prueba es sin límite de tiempo y en general se termina de 30 a 60 minutos.

El 16FP se dirige a estudiantes de preparatoria y adultos. La mayor parte de las normas datan de 1970, lo cual constituye uno de los principales defectos de esta prueba. No obstante, recientemente se obtuvieron normas para la Forma E con poblaciones muy diversas que incluyeron internos en prisiones, pacientes con esquizofrenia, individuos con limitaciones culturales y clientes de rehabilitación física. Sin embargo, la mayoría de los profesionales estaría de acuerdo en que el 16FP es más adecuado para una población "normal" que para una con "trastornos emocionales". Esta prueba también es útil para aplicaciones transculturales (p. ej., Argentero, 1989).

El 16FP se fundamenta en el concepto analítico-factorial de Cattell acerca de la personalidad (Cattell, Eber y Tatsuoka, 1970). Según este modelo, los rasgos superficiales —los aspectos más obvios de la personalidad— surgen de análisis simples de los agrupamientos de respuestas de la prueba. En contraste, los rasgos fuente —los orígenes estables y constantes, pero menos visibles, de la conducta— surgen sólo de análisis factoriales especializados de los rasgos superficiales (Cattell, 1950). En una serie de estudios, Cattell determinó que se necesitan 16 factores o rasgos fuente de la personalidad para explicar la estructura de las respuestas de la prueba, de aquí el nombre de su instrumento.

El 16FP produce un total de 20 índices o atributos de la personalidad. Además de las 16 escalas básicas, se calculan cuatro índices de segundo orden de la personalidad a partir de las sumas lineales ponderadas de los 16 índices anteriores, lo cual produce un total de 20 escalas bipolares. A lo largo de los años, se ha establecido con claridad el significado de las calificaciones extremas en cualquier dirección (cuadro 14–2).

El principal empuje para la aplicación del 16FP es la orientación de carrera, la exploración vocacional y las pruebas de selección laboral. Una razón para la popularidad del instrumento —es el segundo más utilizado después del MMPI/MMPI-2— es que las hojas de respuesta se pueden enviar por correo para calificación electrónica rápida. La mayoría de los profesionales también solicitan un informe narrativo generado por computadora. Una característica atractiva de estos informes es el cúmulo de información que proporcionan. Éstos incluyen una descripción resumida de la personalidad, perfil de calificaciones y resumen de signos clínicos, factores cognitivos y patrones de necesidades.

Una de las principales deficiencias del 16FP es que los 16 atributos de personalidad examinados se basan nada más en 10 a 13 reactivos cada uno. Es inevitable que una prueba con escalas tan cortas como éstas tenga una confiabilidad baja. Por tanto, no es de sorprender que las confiabilidades por mitades de los 16 factores sea tan baja como .54; las correlaciones entre las mismas escalas para las diferentes formas de la prueba se colocan por lo común alrededor de .50 y los coeficientes *test-retest* para las escalas dentro de la misma forma son de .70 y .80 para aplicaciones el mismo día o al siguiente, pero mucho más bajas para intervalos mayores.

La mayor parte de la evidencia en favor del 16FP consiste de demostraciones estadísticas de que los reactivos "pertenecen" a sus respectivas escalas y de que éstas consisten de factores relativamente puros. La evidencia en este sentido es razonablemente alentadora (Cattell, Eber y Tatsuoka, 1970). Además, algunos estudios con el 16FP demuestran que los correlatos objetivos de los resultados de la prueba son consistentes con la teoría. Por ejemplo, Cattell y Nesselroade (1967) estudiaron la semejanza de los perfiles del 16FP de 102 parejas con matrimonios estables y 37 con matrimonios inestables. Estos autores descubrieron que las parejas con matrimonios estables son mucho más similares en el 16FP que aquéllas con matrimonios inestables. El lector observará en el cuadro 14–3 que las correlaciones entre escalas para las parejas con matrimonios estables son casi todas uniformemente positivas —lo cual significa que éstas producen calificacio-

Cuadro 14–2. Los 16 factores de personalidad y 4 índices de segundo orden del 16FP

Nombre del factor	Interpretación de las calificaciones bajas	Interpretación de las calificaciones altas
Calidez	Reservado, lejano, frío, impersonal	Cálido, expresivo, gusta de las personas
Inteligencia	Pensamiento concreto	Pensamiento abstracto, inteligente
Estabilidad emocional	Menos estable a nivel emocional, cambiante	Estable a nivel emocional, calmado, maduro
Dominancia	Sumiso, conformista, plácido	Dominante, asertivo, competitivo
Impulsividad	Serio, prudente, sobrio, taciturno	Entusiasta, alegre, irresponsable
Conformidad	Oportunista, desdeña las reglas	Conformista, perseverante, moralista
Audacia	Tímido, reservado, refrenado	Atrevido, desinhibido, espontáneo
Sensibilidad	Riguroso, confiado en sí mismo	Indulgente, sensible
Suspicacia	Confiado, adaptable	Suspicaz, difícil de engañar, dogmático
Imaginación	Práctico, convencional	Poco práctico, distraído, poco convencional
Astucia	Directo, genuino, sin pretensiones	Calculador, refinado, socialmente alerta
Inseguridad	Confiado, satisfecho consigo mismo, seguro	Se culpa a sí mismo, preocupado, angustiado
Radicalismo	Conservador, se resiste al cambio	Liberal, analítico, innovador
Autosuficiencia	Orientado al grupo, sociable	Ingenioso, autosuficiente
Autodisciplina	Indisciplinado, impulsivo	Compulsivo, socialmente preciso
Tensión	relajado, tranquilo, poco impulso	Frustrado, energético, tenso
Extraversión (Q_I)	Introversión	Extraversión
Ansiedad (Q_{II})	Baja ansiedad	Ansiedad elevada
Apariencia de dureza (Q_{III})	Sensibilidad, emocionalidad	Apariencia de dureza
Independencia (Q_{IV})	Dependencia	Independencia

Fuente: Basado en Wholeben, B. E. (1987). Sixteen Personality Factor Questionnaire. En D. J. Keyser y R. C. Sweerland (eds.). *Test critiques compendium.* Kansas City, MO: Test Corporation of America. También en Cattell, R. B. (1986). *The handbook of the 16 Personality Factor Questionnaire.* Champaign, IL: Institute for Personality and Ability Testing.

nes similares en las escalas del 16FP —mientras que más de la mitad de las correlaciones entre escalas para los matrimonios inestables son negativas— lo cual significa que estas parejas producen con frecuencia calificaciones escalares diferentes en el 16FP. Estos datos sustentan el punto de vista de que la semejanza facilita la estabilidad en el matrimonio.* Más importante aún es que los resultados refuerzan la validez del 16FP al mostrar que los resultados de la prueba tienen implicaciones significativas y anticipatorias en el mundo real.

* Otra posibilidad es que, en los matrimonios estables, las personalidades de los miembros evolucionen a una mayor semejanza con los años. Por supuesto, ambos factores pueden operar de manera simultánea.

Cuadro 14–3. Intercorrelaciones para los factores del 16FP con matrimonios estables e inestables

Factor del 16FP	Correlación para matrimonios estables ($N = 102$)	Correlación para matrimonios inestables ($N = 37$)	Valor p de la diferencia
A Calidez	.16	−.50	<.001
B Inteligencia	.31	.21	ns
C Fortaleza del yo	.32	.05	ns
E Dominancia	.13	.31	ns
F Impulsividad	.23	−.40	<.001
G Conformidad	.33	.19	ns
H Audacia	.23	.12	ns
I Sensibilidad	−.15	−.13	ns
L Suspicacia	.18	−.33	<.01
M Imaginación	.22	−.01	ns
N Astucia	.18	.27	ns
O Inseguridad	.11	.36	ns
Q_1 Radicalismo	.27	.34	ns
Q_2 Autosuficiencia	.15	−.32	<.01
Q_3 Autodisciplina	.27	−.02	ns
Q_4 Tensión	.16	−.11	ns
Q_1 Extraversión	.22	−.30	<.01
Q_2 Ansiedad	.31	.23	ns

Fuente: Adaptado con autorización de Cattell, R. B. y Nesselroade, J.R. (1967). Likeness and completeness theories examined by Sixteen Personality factor measures on stably and unstably married couples. *Journal of Personality and Social Psychology, 7*, 351-361.

CUESTIONARIO DE PERSONALIDAD DE EYSENCK

El Cuestionario de Personalidad de Eysenck (EPQ) es el último en una serie de pruebas diseñadas para medir las principales dimensiones de la personalidad normal y anormal (Eysenck y Eysenck, 1975). Con base en toda una vida de investigación analítico-factorial con el cuestionario y de experimentos de laboratorio sobre aprendizaje y condicionamiento, Eysenck aisló tres dimensiones principales de la personalidad: Psicoticismo (P), Extraversión (E) y Neuroticismo (N). El EPQ consiste de escalas para medir estas dimensiones y también incorpora una escala de Mentira (L) para evaluar la validez de las respuestas de una persona. El EPQ contiene 90 afirmaciones que se responden con "Sí" o "No" y se diseñó para personas de 16 años y mayores. Un EPQ para individuos más jóvenes tiene 81 afirmaciones y es adecuado para niños entre 7 y 15 años de edad.

Los reactivos en la escala P se parecen a los siguientes:

¿Con frecuencia quebrantas las reglas? (V)
¿Te preocuparías si tuvieras deudas? (F)
¿Te expones a riesgos sólo por diversión? (V)

Las calificaciones altas en la escala P indican rasgos agresivos y hostiles, impulsividad, una preferencia por cosas extrañas o poco comunes y defectos de empatía. Es frecuente que los pacientes antisociales y esquizoides obtengan altas calificaciones en esta dimensión. En contraste, las bajas calificaciones en P predicen características más deseables como empatía y

sensibilidad interpersonal. Los reactivos en la escala E se parecen a los siguientes:

¿Te agrada conocer a nuevas personas? (V)
¿Eres silencioso cuando estás con otras personas? (F)
¿Te gustan las grandes emociones? (V)

Las calificaciones altas en la escala E indican una persona llamativa, gregaria, extravertida, amante de las diversiones. Las bajas calificaciones en la escala E representan rasgos de introversión, como la preferencia por la soledad y las actividades tranquilas. Los reactivos en la escala N se parecen a los siguientes:

¿Eres una persona taciturna? (V)
¿Consideras que la vida es aburrida? (V)
¿Es fácil que lastimen tus sentimientos? (V)

La escala N refleja una dimensión de emocionalidad que va de una personalidad nerviosa, desadaptada y excesivamente emocional (altas calificaciones) a una estable y confiada (bajas calificaciones).

La confiabilidad del EPQ es excelente. Por ejemplo, las correlaciones *test-retest* a un mes fueron de .78 (P), .89 (E), .86 (N) y .84 (L). Las consistencias internas se colocaron en el rango de .70 para P y de .80 para las otras tres escalas. La validez de constructo del EPQ también se estableció de manera adecuada a través de docenas de estudios que utilizaron criterios conductuales, emocionales, de aprendizaje, de atención y terapéuticos (revisados en Eysenck y Eysenck, 1976, 1985). Friedman (1987) proporciona una introducción corta pero completa a otras fuentes sobre el EPQ.

Uno de los principales puntos de atención de la investigación con el EPQ se ha colocado en los correlatos empíricos de la **extraversión** y su polo opuesto, la **introversión**. Eysenck y Eysenck (1975) describen al extravertido típico como sigue:

> El extravertido típico es sociable, gusta de las fiestas, tiene muchos amigos, necesita tener personas con quienes hablar y no gusta de leer o estudiar por sí solo. Ansía las emociones, toma riesgos, con frecuencia se destaca, actúa sin pensar y en general es un individuo impulsivo.

Estos autores describen al introvertido típico de la siguiente manera:

> El introvertido típico es el tipo de persona tranquila, retraída, introspectiva que gusta de los libros más que de las personas; es reservado y distante, excepto con amigos íntimos. Tiende a planear de antemano, "mira antes de saltar" y desconfía del impulso del momento.

Eysenck y sus seguidores han vinculado varios factores perceptuales y fisiológicos con la dimensión de extraversión/introversión. Debido a las limitaciones de espacio sólo pueden enumerarse aquí las conclusiones representativas:

- Los introvertidos son más cuidadosos en tareas de vigilancia.
- Los introvertidos tienen mejor desempeño en tareas de detección de señales.
- Los introvertidos son menos tolerantes al dolor, pero más a la privación sensorial.
- A los extravertidos se les puede condicionar con mayor facilidad a estímulos asociados con excitación sexual.
- Los extravertidos tienen mayor necesidad de estimulación externa.

Aiken (1989) resume la investigación adicional sobre los correlatos objetivos de la dimensión de extraversión/introversión del EPQ.

En general, las características técnicas del EPQ son muy fuertes, seguramente más de lo que se encuentra en la mayor parte de los inventarios de autoinforme. La utilidad práctica del instrumento se sustenta en una voluminosa literatura de investigación. No obstante, el EPQ nunca ha atrapado la atención de los psicólogos en EUA, quienes parecen enamorados de los instrumentos multifásicos que producen 10, 20 o 30 calificaciones y no un simple trío de dimensiones básicas.

ESCALAS COMREY DE PERSONALIDAD

Para los profesionales que desean un inventario corto de autoinforme, adecuado para estudian-

tes universitarios y otros adultos, las *Comrey Personality Scales*, CPS (Escalas Comrey de personalidad; Comrey, 1970, 1980), serían la elección adecuada. Como protegido de Guilford, Comrey siguió una estrategia analítico-factorial para desarrollar su prueba de 180 reactivos. Comrey se basó de manera exclusiva en estudiantes universitarios para el desarrollo y estandarización de su prueba, de modo que las CPS son adecuadas para la evaluación de la personalidad en esta subpoblación.

Una virtud especial de las CPS es su brevedad. La prueba consiste de 180 afirmaciones y, por tanto, es de sólo un tercio de la extensión de instrumentos competidores como el MMPI-2. Las ocho escalas de personalidad CPS consisten de 20 reactivos cada una, divididos por igual entre afirmaciones redactadas en dirección positiva y negativa. Otros 20 reactivos se dedican a una verificación de la validez y a evaluar la propensión de respuesta de deseabilidad social.

La siguiente descripción de las escalas CPS se basa en Merenda (1985):

(V) **Verificación de validez.** La puntuación natural esperada es de 8. Sin embargo, cualquier calificación en la escala V que produzca un equivalente en puntuación *T* por debajo de 70 se encuentra todavía dentro del rango normal. Las calificaciones más altas sugieren un protocolo inválido.

(R) **Sesgo de respuesta.** Las calificaciones altas indican una tendencia a responder a las preguntas de manera socialmente deseable, que hace que el individuo parezca una persona "agradable".

(T) **Confianza contra defensividad.** Las altas calificaciones indican una creencia en la honradez básica, confiabilidad y buenas intenciones de otras personas.

(O) **Sentido de orden contra falta de compulsión.** Las calificaciones altas son características de individuos cuidadosos, meticulosos, ordenados y sumamente organizados.

(C) **Conformidad social contra rebeldía.** Los individuos con altas calificaciones aceptan a la sociedad como es, resienten la inconformidad en otros, buscan la aprobación de la sociedad y respetan las leyes.

(A) **Actividad contra falta de energía.** Los individuos con calificaciones altas tienen gran cantidad de energía y resistencia, trabajan de manera intensa y buscan sobresalir.

(S) **Estabilidad emocional contra neuroticismo.** Las personas con calificaciones altas están libres de depresión, son optimistas, relajados, con un estado de ánimo estable y son confiados.

(E) **Extraversión contra introversión.** Los individuos con calificaciones altas conocen personas con facilidad, buscan nuevos amigos, se sienten cómodos con los desconocidos y no sufren de miedo escénico.

(M) **Masculinidad contra feminidad.** Los individuos con calificaciones altas tienden a ser personas bastante rigurosas a quienes no les molesta la sangre, los insectos rastreros, la vulgaridad y que no lloran con facilidad o muestran mucho interés en historias de amor.

(P) **Empatía contra egocentrismo.** Los individuos con calificaciones altas se describen como personas útiles, generosas, compasivas que están interesadas en dedicar su vida al servicio de los demás.

Como reflejo de su cuidadosa derivación analítico-factorial, las escalas CPS poseen excepcionales consistencias internas que van de .91 a .96. Estas conclusiones indican que es más probable que las CPS constituyan una prueba confiable, pero los datos tradicionales de confiabilidad *test-retest* son escasos. Los estudios transculturales con las CPS dan un fuerte apoyo a su validez. Brief y Comrey (1993) informan que los análisis factoriales en muestras rusas, estadounidenses, brasileñas, israelíes, italianas y neozelandesas producen una solución de ocho factores para las respuestas de los reactivos CPS. Otros estudios de validación con estas escalas no son tan directos en su interpretación.

Por un lado, las correlaciones entre las calificaciones de las escalas CPS y los datos biográficos relacionados con la personalidad son muy pequeñas (Comrey y Backer, 1970; Comrey y Schiebel, 1983). Por otro, las calificaciones extremas en las escalas CPS se relacionan en gran medida con trastorno psicológico (Comrey y Schiebel, 1985). Esto es particularmente cierto en cuanto a las bajas calificaciones en Confianza contra defensividad, Actividad contra falta de energía, Estabilidad emocional contra neuroticismo, Extraversión contra introversión y las altas calificaciones en Sentido de orden contra falta de compulsión. En general, los críticos concluyen que las CPS son escalas prometedoras que necesitan actualizar su estandarización además de documentación adicional sobre sus cualidades técnicas.

INVENTARIO NEO DE PERSONALIDAD-REVISADO

El Inventario NEO de Personalidad-Revisado (NEO PI-R) engloba décadas de investigación analítico-factorial con poblaciones adultas, clínicas y normales (Costa y McCrae, 1992). La prueba se basa en el modelo de cinco factores de la personalidad que se describe en el capítulo anterior. Está disponible en dos formas paralelas que consisten de 240 reactivos calificados en una dimensión de cinco puntos. Se utilizan tres reactivos adicionales para verificar la validez. También está disponible una versión más corta, el Inventario NEO de cinco factores (NEO-FFI; Costa y McCrae, 1989). Aquí se limita la discusión al NEO PI-R. La Forma S es para autoinformes, mientras que la Forma R es para observadores externos (p. ej., cónyuge del cliente). El formato de los reactivos consiste de calificaciones en cinco puntos: fuertemente en desacuerdo, en desacuerdo, neutral, de acuerdo, fuertemente de acuerdo. Los reactivos evalúan variables emocionales, interpersonales, experienciales, actitudinales y motivacionales.

Las escalas para los cinco dominios en el NEO PI-R se basan cada una en seis escalas de faceta (rasgo) (cuadro 14–4). La consistencia interna de las escalas es soberbia: de .86 a .95 para las escalas de dominio y .56 a .90 para las escalas de faceta. Los coeficientes de estabilidad van de .51 a .83 en estudios longitudinales de tres a siete años. La evidencia de la validez del NEO PI-R es sustancial y se basa en la correspondencia de las calificaciones entre la persona y su cónyuge, correlaciones con otras pruebas y listas de verificación y en la validez de constructo del modelo de los cinco factores en sí (Costa y McCrae, 1992a; Piedmont y Weinstein, 1993). Trull, Useda, Costa y McCrae, 1995).

El NEO PI-R es una excelente medida de la personalidad normal que es especialmente útil en investigación. La prueba también resulta prometedora como medida de la psicopatología clínica. Por ejemplo, Clarkin, Hull, Cantor y Sanderson (1993) encontraron que los pacientes con un diagnóstico de trastorno limítrofe de la personalidad obtuvieron calificaciones muy altas en Neuroticismo y muy bajas en Conformidad, lo cual concuerda en gran medida con la respuesta de todo clínico hacia estos pacientes desafiantes. Una preocupación menor acerca del instrumento es su falta de escalas sustanciales de validez —sólo tres reactivos evalúan este aspecto. La aplicación del NEO PI-R supone que los individuos son cooperativos y razonablemente sinceros. Ésta es en general una suposición segura en entornos de investigación, pero puede no resultar cierta en ambientes forenses, de selección de personal y psiquiátricos.

INVENTARIOS CON CODIFICACIÓN DE CRITERIO

Los inventarios de autoinforme que se revisarán por último engloban una estrategia de desarrollo de una prueba con codificación de criterio. En un enfoque de codificación de criterio, los reactivos de una prueba se asignan a una escala particular si, y sólo si, discriminan entre un grupo de criterio bien definido y uno control relacionado. Por ejemplo, para diseñar

Cuadro 14–4. Escalas de dominio y facetas (rasgos) del NEO PI-R

Dominios	Facetas	
Neuroticismo	Ansiedad	Timidez
	Hostilidad con enojo	Impulsividad
	Depresión	Vulnerabilidad
Extraversión	Calidez	Actividad
	Sociabilidad	Búsqueda de emociones
	Asertividad	Emociones positivas
Apertura a la experiencia	Fantasía	Acciones
	Estética	Ideas
	Sentimientos	Valores
Conformidad	Confianza	Sumisión
	Franqueza	Modestia
	Altruismo	Indulgencia
Escrupulosidad	Competencia	Esfuerzo dirigido al logro
	Orden	Autodisciplina
	Diligencia	Deliberación

una escala de autoinforme para la depresión, los reactivos que las personas deprimidas confirman con una frecuencia significativamente mayor (o menor) que los controles normales se asignarían a la escala de depresión y se codificarían en la dirección apropiada. Podría utilizarse un enfoque similar para otros constructos de interés para el clínico, como la esquizofrenia, reacción de ansiedad y similares. Véase que el autor de la prueba no consulta ninguna teoría de la esquizofrenia, la depresión o la reacción de ansiedad para determinar cuáles reactivos pertenecen a las diferentes escalas. La esencia del procedimiento con codificación de criterio es, por así decir, dejar que los reactivos caigan donde sea.*

* En este caso se están pasando por alto ciertas características complejas. Algunos reactivos que reflejan psicopatología general podrían discriminar a **todos** los grupos de contraste con respecto al grupo control. El autor de la prueba podría descartar éstos en favor de reactivos que discriminen **de manera diferencial** en el caso de un grupo de contraste pero no en los demás.

INVENTARIO MULTIFÁSICO DE LA PERSONALIDAD DE MINNESOTA-2 (MMPI-2)

Publicado inicialmente en 1943, el MMPI era un inventario de personalidad con 566 reactivos de cierto/falso diseñado en un principio como auxiliar para el diagnóstico psiquiátrico (Hathaway y McKinley, 1940, 1942, 1943; McKinley y Hathaway, 1940, 1944; McKinley, Hathaway y Meehl, 1948). Los autores de la prueba siguieron un estricto enfoque de codificación de criterio para la construcción de las escalas del MMPI. Las escalas clínicas se desarrollaron al contrastar las respuestas de grupos cuidadosamente definidos de pacientes psiquiátricos (*N* promedio de cerca de 50 personas) con las respuestas de 724 personas control. El resultado fue una notable prueba útil tanto en la evaluación psiquiátrica como en la descripción de la personalidad normal. Después de unos cuantos años, el MMPI se convirtió en la prueba de personalidad de uso más amplio en EUA.

Al principio, el MMPI envejeció con gracia; los profesionales toleraron lo que parecían

ser defectos menores. Pero a medida que la prueba alcanzó la mediana edad, se volvió cada vez más obvia la necesidad de un rejuvenecimiento. El problema más grave residía en el grupo control original, que consistió principalmente de familiares y visitantes de los pacientes médicos en el Hospital de la Universidad de Minnesota. La elección estrecha de personas control, examinadas principalmente durante el decenio de 1930, resultó ser una pertinaz fuente de críticas hacia el instrumento. Todas las personas control eran de raza blanca y la mayoría eran jóvenes (edad promedio 35 años), casadas y provenientes de un pequeño pueblo o del área rural. Ésta fue una muestra de conveniencia que era significativamente no representativa de la población en general.

El contenido de los reactivos del MMPI también hizo surgir preocupaciones (Graham, 1993). Varios de ellos utilizaban una terminología arcaica y obsoleta, como "tirar el pañuelo" (un juego de salón del decenio de 1930), polvos para dormir (píldoras para dormir) y tranvías (autobuses eléctricos). Otros empleaban un lenguaje sexista. Los individuos examinados pensaban que algunos reactivos eran inaceptables, en especial aquellos que se relacionaban con creencias religiosas cristianas. Estos reactivos fueron causa de demandas legales ocasionales que alegaban una invasión de la vida privada. Por último, unos cuantos reactivos que trataban sobre funciones fisiológicas y conducta sexual eran francamente ofensivos.

Desde el punto de vista de la medición, un problema más grave en el contenido de los reactivos era la omisión. El conjunto original de reactivos del MMPI no tenía la suficiente amplitud como para evaluar muchas características importantes, como tendencias suicidas, abuso de estupefacientes y conductas relacionadas con el tratamiento. Un motivo adicional para la revisión del MMPI consistió en aumentar el rango de cobertura de los reactivos.

El MMPI-2 se publicó en 1989 después de casi una década de revisión y reestandarización. El nuevo y mejorado MMPI-2 incorpora una muestra normativa contemporánea de 2 600 individuos que son aproximadamente representativos de la población general en cuanto a las variables demográficas principales (localización geográfica, raza, edad, nivel ocupacional e ingresos). Aunque los niveles superiores de educación tienen una representación excesiva, la muestra normativa del MMPI-2 sigue siendo una notable mejoría con respecto a la del MMPI. Se ha mejorado de manera significativa la reserva de reactivos al revisar aquéllos obsoletos, eliminar los ofensivos y añadir nuevos reactivos para extender la cobertura del contenido.

El MMPI-2 representa una mejoría significativa con respecto al MMPI, pero conserva una continuidad sustancial con su estimado antecesor. Los autores de la prueba conservaron los mismos títulos y objetivos de medición para las escalas tradicionales clínicas y de validez. La reestandarización proporciona una mejor calibración para las elevaciones de escala, una mejoría que era muy necesaria (Tellegen y Ben-Porath, 1992). Aunque se volvieron a redactar docenas de reactivos, estas revisiones son de apariencia y no afectan las características psicométricas de la prueba (Ben-Porath y Butcher, 1989). De hecho, cuando muestras grandes de personas responden al MMPI y al MMPI-2, las calificaciones para las escalas individuales clínicas y de validez se correlacionan por lo común en cerca de .99.

El MMPI-2 consiste de 567 reactivos cuidadosamente diseñados para evaluar una amplia gama de preocupaciones. Se le pide al examinando que marque "Cierto" o "Falso" para cada afirmación, según se le aplique. La mayor parte de los reactivos son autorreferenciales. Los reactivos engloban los siguientes temas generales (Dahlstrom, Welsh y Dahlstrom, 1972; Graham, 1993):

Preocupaciones de salud.
Síntomas neurológicos.
Problemas familiares.
Relaciones matrimoniales.
Problemas laborales.
Síntomas depresivos y maniacos.
Estados obsesivos y compulsivos.
Delirios y alucinaciones.
Ansiedad y fobias.

Control del enojo.
Prácticas antisociales.
Abuso de sustancias prohibidas y alcohol.
Autoestima.
Incomodidad social.
Conducta tipo A.
Masculinidad-feminidad.
Indicadores negativos de tratamiento.
Validez de respuesta; por ejemplo, virtudes improbables.

El MMPI requiere un nivel de lectura de sexto grado y la mayoría de los individuos termina de responderlo en 1 a 1.5 horas.

Las escalas originales del MMPI se desarrollaron al contrastar las respuestas a los reactivos de grupos cuidadosamente definidos de pacientes psiquiátricos (*N* promedio de 50) con aquéllas de aproximadamente 700 personas control. Los grupos de pacientes psiquiátricos incluyeron las siguientes categorías diagnósticas: hipocondriasis, depresión, histeria, psicopatía, homosexualidad masculina, paranoia, psicastenia*, esquizofrenia y fases iniciales de manía (hipomanía). Además, se utilizaron muestras de estudiantes universitarios socialmente introvertidos y extravertidos para construir una escala de introversión social. El MMPI-2 conserva las escalas clínicas básicas, con unos cuantos cambios menores de eliminación y revisión. Ben-Porath y Butcher (1989) investigaron las características de los reactivos que volvieron a redactarse en el MMPI-2 y descubrieron que son psicométricamente equivalentes a los originales.

El MMPI-2 puede calificarse para cuatro escalas de validez, 10 escalas clínicas estándar y docenas de escalas complementarias. En la práctica, los clínicos le dan mayor énfasis a las escalas clínicas estándar y a las de validez. Las escalas complementarias son sólo eso —un complemento. Proporcionan información útil para afinar la interpretación de las escalas clínicas y de validez tradicionales. Las puntuaciones naturales en las escalas del MMPI-2 se convierten a puntuaciones *T*, con una media de 50 y desviación estándar de 10. Las calificaciones que superan la puntuación *T* de 65 merecen consideración especial. Éstas son estadísticamente poco comunes dentro de la población general y pueden significar la presencia de sintomatología psiquiátrica. Aquí nos concentraremos en las escalas tradicionales, comenzando con una revisión de las cuatro escalas de validez conocidas como No puedo decir (?), L, F y K.

La calificación de frases omitidas es simplemente el número total de reactivos omitidos o con doble respuesta en el llenado de la hoja de respuestas. Las instrucciones para la prueba estimulan a las personas examinadas a responder todos los reactivos, pero ocurrirán omisiones o dobles respuestas. No obstante, esto es poco común —el número modal de reactivos omitidos es de cero (Tamkin y Scherer, 1957). La omisión de hasta 10 reactivos parece tener poco efecto sobre los resultados globales de la prueba —uno de los beneficios de tener un conjunto tan grande de afirmaciones en el MMPI-2. Una calificación muy alta en esta escala puede indicar un problema de lectura, oposición a la autoridad, defensividad o indecisión producida por depresión.

La Escala L se compone de 15 reactivos que se califican todos en la dirección de falso. Al responder "falso" a los reactivos de la Escala L, la persona examinada afirma que posee un grado de virtudes personales que se observa en pocas ocasiones en nuestra cultura (p. ej., nunca se enoja, siente agrado por todos, nunca miente, lee todos los editoriales del periódico y preferiría perder a ganar). La Escala L se diseñó para identificar una actitud general deliberadamente evasiva al realizar la prueba. Una calificación alta en la Escala L indica que el individuo no sólo es defensivo, sino que también lo hace de manera inocente. Las personas con algún grado de complejidad psicológica pueden adoptar una actitud defensiva al responder la prueba y de todos modos obtener una calificación dentro del rango normal en la Escala L.

La Escala F consiste de 60 reactivos que los individuos normales responden en la direc-

* Este término diagnóstico arcaico es bastante similar a lo que ahora se denominaría trastorno obsesivo-compulsivo.

ción calificada en no más de 10% de las ocasiones. Éstos reflejan un amplio espectro de desadaptación grave, incluyendo pensamientos peculiares, apatía y enajenación social. Aunque los reactivos de la Escala F parecen indicar patología psiquiátrica, en pocas ocasiones los pacientes responden a ellos en la dirección esperada. Menos de 50% de estos reactivos aparecen en las escalas clínicas. No obstante, muchas personas con trastornos psiquiátricos importantes sí llegan a producir calificaciones elevadas, en el rango de $T = 70$ u 80 en la Escala F. Por otro lado, las calificaciones excepcionalmente elevadas sugieren hipótesis adicionales: capacidad de lectura insuficiente, respuestas aleatorias o de falta de cooperación, un intento motivado de "parecer malo" en la prueba o un exagerado "llamado de ayuda" en el caso de un cliente angustiado.

La Escala K se diseñó para ayudar a detectar una forma sutil de defensividad. La escala con 30 reactivos se compone, en parte, de 22 que diferencian los perfiles normales producidos por pacientes psiquiátricos hospitalizados que muestran defensividad de aquellos que producen los controles normales. Además se añadieron ocho reactivos que mejoran la discriminación de síntomas depresivos y esquizofrénicos (McKinley, Hathaway y Meehl, 1948). Una calificación elevada en la Escala K puede indicar una actitud defensiva al responder la prueba. Las elevaciones dentro del rango normal en la Escala K sugieren buena fortaleza del yo —la presencia de defensas psicológicas útiles que permiten que la persona funcione bien a pesar de conflictos internos.

El uso combinado de las escalas F y K puede ser útil para la detección de perfiles del MMPI-2 donde existe falsificación o simulación. En un estudio, 81% de los perfiles en los que se simulaba "ser bueno" se identificaron a través de la sencilla regla de decisión (utilizando las puntuaciones naturales) de $F\text{-}K < 12$, mientras que 87% de los perfiles en los que se simulaba "ser malo" se identificaron por una simple regla de decisión (utilizando puntuaciones naturales) de $F\text{-}K > 7$ (Bagby, Rogers, Buis y Kalemba, 1994). Los estudios en los que se entrena a las personas para falsear el MMPI-2 hacer surgir preocupaciones éticas perturbadoras (Ben-Porath, 1994; Wetter, Baer, Berry y Reynolds, 1994).

Varias escalas clínicas tienen una "corrección K" para mejorar su poder de discriminación. La fundamentación para esta práctica es que las elevaciones de K indican una reducción artificial de las calificaciones en estas escalas clínicas. Por tanto, partes de la puntuación natural de K se suman a las calificaciones en estas escalas clínicas antes de calcular las puntuaciones *T*. Las escalas con corrección K, que se analizan más adelante, incluyen Hipocondriasis, Desviación psicopática, Psicastenia, Esquizofrenia e Hipomanía. El que en realidad la corrección K mejore al MMPI-2 es debatible, pero los editores de la prueba continuaron la tradición del MMPI en beneficio de la continuidad. También se encuentran disponibles normas independientes para transformaciones de calificaciones de escala sin corrección K.

Además de las escalas de validez, en el MMPI-2 siempre se califica para 10 escalas clínicas. Con excepción de Introversión social, dichas escalas se construyeron a la manera común de la codificación de criterio al contrastar las respuestas de pacientes clínicos con las de controles normales. Como se señaló antes, Introversión social se desarrolló a través de contrastar las respuestas de estudiantes universitarios con alta y baja introversión social. En el cuadro 14–5 se resumen las 10 escalas clínicas y las interpretaciones comunes para las calificaciones elevadas.

También pueden obtenerse calificaciones para docenas de escalas complementarias en el MMPI-2. Algunas de éstas se basan en la identificación racional de conjuntos de síntomas y en la purificación subsecuente de las escalas a través de medios empíricos. De esta manera se desarrollaron en el MMPI-2 15 interesantes Escalas de contenido (Butcher, Graham, Williams y Ben-Porath, 1990). Fueron investigadores independientes quienes desarrollaron muchas de las escalas complementarias; éstas varían ampliamente en cuanto a calidad. En la práctica,

Cuadro 14–5. Las 10 escalas clínicas del Inventario multifásico de la personalidad de Minnesota-2

Núm. de escala y abreviatura	Nombre de la escala	Corrección K	Interpretación típica de la elevación
1 Hi	Hipocondriasis	.5K	Preocupación física excesiva
2 D	Depresión		Sentimientos de tristeza, desesperanza
3 Hy	Histeria		Inmadurez, uso de represión, negación
4 Dp	Desviación psicopática	.4K	Conflictos con autoridad, impulsividad
5 Mf	Masculinidad-feminidad		Intereses masculinos [mujeres], Intereses femeninos [varones]
6 Pa	Paranoia		Suspicacia, hostilidad
7 Pt	Psicastenia	1K	Ansiedad y pensamientos obsesivos
8 Es	Esquizofrenia	1K	Distanciamiento, procesos extraños de pensamiento
9 Ma	Hipomanía	.2K	Gran energía, posible agitación
0 Is	Introversión social		Timidez e introversión

sólo cerca de 30 de las escalas adicionales se califican de modo rutinario. Ejemplos de las escalas complementarias incluyen Ansiedad, Represión, Fortaleza del yo y la Escala MacAndrew de alcoholismo–Revisada. Ansiedad (A) y Represión (R) son los primeros dos factores principales que siempre surgen del análisis factorial de las respuestas del MMPI-2. Una interesante escala complementaria es la Escala de Fortaleza del yo (Es) de Barron, cuya pretensión es anticipar la respuesta positiva a la psicoterapia. No obstante, no todos los estudios confirman este uso de la escala (Graham, 1987). La Escala MacAndrew de alcoholismo -Revisada (MAC-R; MacAndrew, 1965) es un índice útil del abuso de alcohol y de otras sustancias. La MAC-R no sólo es eficaz en la evaluación del alcoholismo, sino que también es útil para identificar a bebedores extremos y a individuos con farmacodependencia (Wolf, Schubert, Patterson, Grande, Pendleton, 1990). En este caso es imposible revisar todas las escalas complementarias útiles. El lector interesado debería consultar a Butcher y Williams (1992) y Graham (1993).

INTERPRETACIÓN DEL MMPI-2

La interpretación de un perfil del MMPI-2 puede seguir dos vías diferentes: escala por escala o a través de la configuración. En el enfoque más sencillo posible, aquél de escala por escala, el examinador determina la validez de la prueba, como se analiza antes, al inspeccionar las cuatro escalas de validez. Si la prueba parece tener una validez razonable según estos criterios, el examinador consulta un libro pertinente y procede escala por escala para producir una serie de hipótesis. Por ejemplo, Lachar (1974) ha obtenido el significado de diversas elevaciones en la escala Pa o Paranoia como sigue:

- $T = 27$ a 44, el individuo examinado puede ser testarudo, irritable o difícil.
- $T = 45$ a 59 no existe sensibilidad indebida y hay una adecuada consideración hacia los demás.
- $T = 60$ a 69 aumento de probabilidad de rigidez y sensibilidad excesiva.
- $T = 70$ a 79 rígido, irritable, proyecta culpa y hostilidad.

T = 79 a 100 pueden estar presentes rasgos francamente paranoides y delirantes.

El enfoque de configuración para la interpretación del MMPI-2 es un tanto más complicado y consiste en clasificar el perfil como perteneciente a una u otra clave de perfil definida de manera aproximada y que se ha estudiado extensamente. En general, las claves de perfil se definen por una combinación de elevación (dos o más escalas clínicas elevadas por encima de cierto criterio) y definición (dos o más escalas clínicas que se destacan de manera clara de las demás). Por ejemplo, en su manifestación más evidente, la clave de perfil 4-9 puede definirse por un perfil válido en el que la escala 4 (Desviación psicopática) y la escala 9 (Hipomanía) son las elevaciones más altas, que superan ambas a una T de 65 (elevación) y ambas exceden a la siguiente escala clínica más alta en, cuando menos, 5 puntos de puntuación T (definición). Ésta es la manera en que Graham (1993) describe a las personas que encajan en esta clave de perfil:

> La característica más sobresaliente de los individuos 49/94 es un notable menosprecio hacia las normas y valores sociales. Con frecuencia tienen problemas con las autoridades debido a su conducta antisocial. Tienen una conciencia poco desarrollada, moral disipada y valores éticos cambiantes. Entre las dificultades en las que pueden verse implicados se encuentra el alcoholismo, las peleas, problemas matrimoniales, *acting out* sexual y un amplio conjunto de actos delictivos. Ésta es una clave de perfil común entre personas que abusan del alcohol y otras sustancias.

El diagnóstico más probable para dichos individuos es el trastorno de personalidad antisocial.

Se debe mencionar brevemente que se encuentran disponibles varios sistemas de interpretación por computadora para el MMPI y el MMPI-2 (Fowler, 1985; Butcher, 1987). El *Minnesota Report*™ (Butcher, 1993) es el mejor. Este sistema genera un informe muy cauto y metódico de 16 páginas que incluye el análisis de la validez del perfil, patrones sintomáticos, relaciones interpersonales, consideraciones diagnósticas y de tratamiento. El *Minnesota Report*™ también proporciona una variedad de figuras y tablas para ilustrar los resultados de prueba.

La calidad de los informes narrativos por computadora para el MMPI-2 es generalmente buena, pero el lector debería percatarse de que son seres humanos falibles quienes escriben estos tipos de programas. Existe el peligro de que los informes de la prueba generados por computadora sean erróneos. Lo que es más, por unos cuantos cientos de dólares se pueden comprar en disquete de microcomputadora algunos sistemas de interpretación que son poco dignos de confianza. Esto aumenta el riesgo de que personas poco calificadas den un mal uso a las interpretaciones de la prueba por medio de computadora. En el capítulo final de esta obra se analizan las dificultades de la interpretación por computadora.

PROPIEDADES TÉCNICAS DEL MMPI-2

Desde el punto de vista de los criterios psicométricos tradicionales, el MMPI-2 presenta una imagen confusa. Los datos de confiabilidad son positivos en general, con coeficientes medianos de consistencia interna (alfa) que por lo común se colocan en los rangos de .70 y .80, pero que llegan a niveles muy bajos, dentro del rango de .30, para algunas escalas en algunas muestras. Los coeficientes *test-retest* a una semana van de la parte alta del rango de .50 a la parte baja del de .90, con una mediana dentro del rango de .80 (Butcher, Dahlstrom, Graham, Tellegen y Kaemmer, 1989). Éstas son buenas cifras si se considera que algunos atributos —como aquéllos medidos con la escala de Depresión— cambian con tal rapidez que la metodología *test-retest* tiene una idoneidad cuestionable.

Una de las principales deficiencias del MMPI-2 es que las intercorrelaciones entre las escalas clínicas son extremadamente altas. Por ejemplo, en el caso de las escalas 7 y 8 (Psicastenia y Esquizofrenia), la correlación se encuentra por lo común en el rango de .70. En parte, esto refleja la superposición de los reactivos entre las escalas del MMPI —las escalas 7 y 8 comparten 17 reactivos. Pero también es cierto que

el enfoque de codificación de criterio no es muy adecuado para el desarrollo de medidas independientes. Una alta intercorrelación de las escalas básicas es uno de los precios a pagar por utilizar esta estrategia de desarrollo de pruebas.

La validez del MMPI-2 es difícil de resumir, debido al enorme volumen de investigación sobre este instrumento y su antecesor, el MMPI. Para 1975 se habían llevado a cabo más de 6 000 estudios con el MMPI (Dahlstrom, Welsh y Dahlstrom, 1975). Por supuesto, miles de estudios adicionales se han publicado desde entonces. Graham (1993) proporciona una breve pero interesante reseña de los estudios de validez sobre el MMPI/MMPI-2. Este autor señala que los coeficientes promedio de validez para los estudios sobre el MMPI entre 1970 y 1981 era un saludable .46. También destaca el patrón de confirmación que constituyen los correlatos externos a la prueba en docenas de estudios de grupos identificados de pacientes. Por último, la investigación indica que el MMPI-2 es muy comparable al MMPI, para el cual se ha recopilado un cuerpo sustancial de datos de validez (Hargrave, Hiatt, Ogard y Karr, 1994; Harrell, Honaker y Parnell, 1992). Por último, los estudios sobre sesgo que comparan los resultados en el MMPI-2 de clientes caucásicos y afroestadounidenses indican que existen ligeras diferencias raciales en los perfiles promedio. No obstante, estas diferencias reflejan de manera válida el funcionamiento emocional; es decir, el MMPI-2 no tiene un sesgo racial (McNulty, Graham, Ben-Porath y Stein, 1997).

Se ha mostrado que el MMPI y el MMPI-2 son valiosos para una amplia gama de problemas diagnósticos, incluyendo la evaluación de los trastornos antisociales, limítrofes y narcisistas de la personalidad (Castlebury, Hilsentoth, Handler y Durham, 1997), la valoración de los antecedentes de abuso sexual y la orientación sexual en mujeres (Griffith, Myers, Cusick y Tankersley, 1997), la predicción del resultado de cirugía para lumbalgia (Uomoto, Turner y Herron, 1988) y la clasificación empírica de personas acusadas de homicidio (Kalichman, 1988), por citar sólo unos cuantos ejemplos. Es probable que durante muchos años por venir el MMPI-2 conserve su estatus como el principal instrumento para la evaluación de la psicopatología en la adultez.

INVENTARIO PSICOLÓGICO DE CALIFORNIA (CPI)

Publicado originalmente en 1957, el CPI es un instrumento similar al MMPI que se revisó de manera reciente y se diseñó de modo expreso para medir las dimensiones de la personalidad normal (Gough, 1987; McAllister, 1986). La prueba consiste de 462 reactivos de cierto/falso, incluyendo cerca de 200 que se tomaron prestados en forma directa del MMPI. El CPI revisado produce calificaciones en 20 escalas que incluyen tres medidas diseñadas para evaluar la actitud al responder la prueba. Estas escalas de validez son Sentido de bienestar (Sb), obtenida de personas normales a quienes se pidió que se "fingieran malos"; Buena impresión (Bi), extraída de personas normales a quienes se pidió que se "fingieran buenos"; y Aspectos comunes (Cm), que consiste de los reactivos refrendados por 95% de los individuos normales. Así, Sb, Bi y Cm están diseñadas, respectivamente, para detectar a los individuos que se fingen malos, buenos o responden de manera aleatoria.

Las escalas clínicas se basan en conceptos "populares" acerca de la personalidad; miden dimensiones de ésta que son significativas y fácilmente reconocibles para legos y psicólogos por igual. Para 13 de las 17 escalas clínicas, Gough utilizó un enfoque de codificación de criterio para el desarrollo de la prueba. Se formaron grupos extremos de participantes (principalmente estudiantes universitarios) según criterios relacionados con la escala como calificaciones escolares, sociabilidad y participación en actividades extracurriculares. Después se contrastaron las frecuencias de confirmación de los reactivos para revelar las mejores afirmaciones para cada escala. Las cuatro escalas restantes se construyeron sobre una base racional respaldada por índices de consistencia interna. Las escalas CPI se describen en el cuadro 14–6.

Como reflejo del cuidado con el que se construyeron las escalas, los datos de confiabilidad

Cuadro 14–6. Descripción breve de las Escalas del Inventario psicológico de California

Do	Dominancia	Dominante, persistente, buena capacidad de liderazgo
Ce	Capacidad de estatus	Cualidades personales que subyacen y conducen al estatus
Sb	Sociabilidad	Temperamento comunicativo, sociable, participativo
Ps	Presencia social	Aplomo, espontaneidad y confianza en sí mismo en situaciones sociales
Aa	Autoaceptación	Autoaceptación y sentido de valía personal
In	Independencia	Fuerte sentido de independencia personal, no se le influye con facilidad
Sb	Sentido de bienestar	No se preocupa o se queja, libre de dudas sobre sí mismo
Em	Empatía	Buena capacidad para empatizar con otras personas
Re	Responsabilidad	Escrupuloso, responsable y confiable
So	Socialización	Fuerte madurez social y elevada integridad
Cm	Autocontrol	Buen autocontrol, libertad de la impulsividad y del egocentrismo
To	Tolerancia	Permisivo, aceptante y poco crítico de las creencias sociales
Bi	Buena impresión	Preocupado por crear una buena impresión
Ac	Aspectos comunes	Patrón de respuesta válida y cuidadosa
Lc	Logro a través de conformidad	Buenos logros en ambientes donde es necesaria la conformidad
Li	Logro a través de independencia	Buenos logros en ambientes donde es necesaria la independencia
Ei	Eficiencia intelectual	Alto grado de eficiencia personal e intelectual
Cp	Comprensión psicológica	Interesado en las necesidades internas, motivos y experiencias de otros, además de responder a ellas
Fx	Flexibilidad	Flexible y adaptable en pensamiento y conducta social
Fe	Feminidad	Alto grado de intereses femeninos

Fuente: Basado en Gough, H. G. (1987). *California Psychological Inventory manual.* Palo Alto, CA: Consulting Psychologists Press. También en Megargee, E. (1972). *The California Psychological Inventory handbook.* San Francisco: Jossey-Bass.

para el CPI son bastante satisfactorios. La mayor parte de las correlaciones *test-retest* entre escalas se encuentran en el nivel de .80 a .90, con un rango de aproximadamente los .50 hasta la parte baja de la zona de los .90, dependiendo de la muestra y del intervalo entre aplicaciones. El CPI también tiene una buena estandarización; las calificaciones se basan en una muestra normativa de 6 000 varones y 7 000 mujeres de edades, clases sociales y regiones geográficas muy variadas. Todas las puntuaciones escalares se presentan como puntuaciones *T*, con una media de 50 y una desviación estándar de 10.

Las escalas CPI se sometieron a una validación cruzada para determinar su capacidad para discriminar a nuevas muestras de individuos calificados en la dimensión pertinente. Estos coeficientes de validez son bastante confusos: algunos valores son aceptables, pero otras correlaciones de criterio son muy bajas. Por ejemplo, es común que las calificaciones en la escala de Flexibilidad se correlacionen sólo en cerca de .2 con las calificaciones de flexibilidad de parte del personal (Domino, 1987).

El CPI es heredero de una larga historia de investigación empírica que corrobora varios correlatos reales para perfiles distintivos de la

prueba. Debido a las limitaciones de espacio, sólo se pueden enumerar varias áreas destacadas en las que se ha confirmado de manera empírica el valor de la prueba. El CPI es útil como auxiliar para predecir lo siguiente:

- Rendimiento en preparatoria y universidad.
- Efectividad de estudiantes-maestros.
- Promedio general de calificaciones en facultades de medicina.
- Eficacia del personal policiaco y militar.
- Liderazgo y éxito ejecutivo.

El CPI tiene particular efectividad para identificar adolescentes o adultos con una vida de delincuencia o criminalidad (Gough y Bradley, 1992b). El lector puede encontrar mayores detalles sobre los correlatos empíricos objetivos de los perfiles del CPI en Groth-Marnat (1990) y Hargrave y Hiatt (1989).

INVENTARIO MULTIAXIAL CLÍNICO DE MILLON-III (MCMI-III)

El MCMI-III es un inventario de la personalidad diseñado para los mismos propósitos que el MMPI-2; es decir, para proporcionar información útil para el diagnóstico psiquiátrico (Millon, 1983, 1987, 1994). El MCMI-III tiene cuando menos dos ventajas sobre el MMPI-2: primero es mucho más corto (175 reactivos de cierto-falso) y, por tanto, es más aceptable para las canalizaciones clínicas; segundo, está planeado y organizado para identificar patrones clínicos de una manera que es compatible con el Manual diagnóstico y estadístico (DSM-IV) de la *American Psychiatric Association*.

El MCMI-III es una prueba con bases teóricas muy fuertes, al incorporar la elaborada formulación teórica de Millon sobre la naturaleza de la psicopatología y de los trastornos de personalidad (Millon, 1969, 1981, 1986; Millon y Davis, 1996). La prueba incluye 27 escalas que se listan en el cuadro 14–7. Las primeras 11 miden estilos o rasgos de personalidad como el narcisismo y las tendencias antisociales; las siguientes tres evalúan una patología más grave de la personalidad (trastornos esquizotípico, limítrofe y paranoide); las siguientes siete escalas evalúan síndromes clínicos como ansiedad y depresión; las siguientes tres escalas evalúan síndromes clínicos más graves como trastornos del pensamiento; y las últimas tres son índices de validez (estilo de respuesta). Las calificaciones en estas escalas (Divulgación, Deseabilidad

Cuadro 14–7. Escalas del Inventario Multiaxial Clínico de Millon–III

Patrones clínicos de personalidad		**Síndromes clínicos**	
1	Esquizoide	A	Ansiedad
2A	Evitante	H	Somatoforme
2B	Depresivo	N	Bipolar: Maniaco
3	Dependiente	D	Distimia
4	Histriónico	B	Dependencia del alcohol
5	Narcisista	R	Trastorno por estrés postraumático
6A	Antisocial		
6B	Agresivo (Sádico)	**Síndromes graves**	
7	Compulsivo	SS	Trastorno del pensamiento
8A	Pasivo-Agresivo (Negativista)	CC	Depresión mayor
8B	Autoderrotante	PP	Trastorno delirante
Patología grave de la personalidad		**Índices de validez (Modificadores)**	
S	Esquizotípico	X	Divulgación
C	Limítrofe	Y	Deseabilidad
P	Paranoide	Z	Degradación

y Degradación), se utilizan para ajustar las otras puntuaciones de escala hacia arriba o hacia abajo, con base en la defensividad o exageración de los síntomas, respectivamente.

El desarrollo de la escala para MCMI-III y sus precursores fue cuidadoso y metódico. En este caso sólo se puede describir un bosquejo amplio de este proceso, donde se seleccionaron 175 afirmaciones de 3 500 reactivos originales durante tres etapas de desarrollo de la prueba: una etapa téorica-de sustentación (redacción de reactivos con base en la teoría), una etapa interna-estructural (correlaciones reactivo-escala) y una etapa de criterio externo (contrastación de grupos diagnósticos con el grupo de referencia). Una característica especial de la última etapa fue que Millon empleó como grupo de referencia a pacientes psiquiátricos en general, más que controles normales. El propósito de esta estrategia era aumentar la capacidad de las escalas del MCMI para diferenciar entre sí a los grupos diagnósticos específicos. Por desgracia, un efecto secundario de este enfoque particular de codificación de criterio fue un grado bastante sustancial de superposición de reactivos en las escalas clínicas. Millon había planeado y esperado la superposición de reactivos, pero es probable que no haya anticipado que algunos pares de escalas en el MCMI compartirían la mayoría de sus reactivos en común. El posterior refinamiento de la escala en la segunda y tercera ediciones eliminó parte de esta superposición. El instrumento revisado también incorpora un procedimiento de ponderación de reactivos. En este enfoque, las preguntas individuales recibieron una ponderación de 2 o 1 para reflejar su importancia en la discriminación del prototipo de cada escala. El enfoque de ponderación de reactivos ha recibido la crítica de ser innecesario y poco flexible (Streiner, Goldberg y Miller, 1993).

La muestra normativa del MCMI-III consistió de cerca de 1 000 varones y mujeres pacientes, en todo EUA. Éste es un enfoque poco común y polémico para la obtención de una muestra normativa. Es más típico que se utilice el muestreo proporcionado de la población de individuos razonablemente normales. Millon ofrece la discutible justificación de que una muestra de pacientes es adecuada como muestra normativa debido a que se consultaron las tasas base (en la población general) de trastornos específicos, clínicos y de personalidad, para calibrar las puntuaciones de corte en las escalas individuales (Millon y Davis, 1996). Pero este enfoque es complejo, experimental y difícil de comprender. La confiabilidad de las escalas individuales es bastante buena: los coeficientes de consistencia interna promedian de .82 a .90 y los coeficientes *test-retest* para una semana van de .81 a .87. La sustentación para la validez del MCMI-III es confusa (Haladyna, 1992; Piersma y Boes, 1997).

Es poco probable que el MCMI-III reemplace al MMPI-2, que es más adecuado para el diagnóstico de síndromes clínicos agudos. Sin embargo, el MCMI-III resulta prometedor para el diagnóstico de trastornos de personalidad y, por tanto, puede complementar en este sentido al MMPI-2 (Antoni, 1993). No obstante, varios estudios independientes actuales han puesto en duda el valor del MCMI-III y en especial de sus precursores en la determinación de diagnósticos dentro del esquema del Manual Diagnóstico y Estadístico de los Trastornos Mentales (Smith, Carroll y Fuller, 1988; Patrick, 1988; McCann y Suess, 1988). También, Choca, Shanley, Peterson y van Denburg (1990) encontraron evidencia de sesgo racial contra los afroestadounidenses en el MCMI. Reynolds (1992) capta lo que debe ser la opinión compartida de muchos clínicos cuando describe al MCM como "una joya conceptual que a nivel psicométrico se encuentra entre una pesadilla y un enigma". Para otras críticas, véanse Dana y Cantrell (1988), y Overholser (1990). Craig (1993) ha conjuntado una serie de artículos que en su mayor parte apoyan al MCMI.

INVENTARIO DE PERSONALIDAD PARA NIÑOS

El *Personality Inventory for Children*; PIC (Inventario de Personalidad para Niños) utiliza el

enfoque metodológico general incorporado en el MMPI y se diseñó para proporcionar descripciones clínicas pertinentes de la conducta infantil y de las características familiares (Lachar, 1982; Wirt y Lachar, 1981; Wirt, Lachar, Klinedinst y Seat, 1984). El PIC puede aplicarse a niños y adolescentes entre los 3 y 16 años de edad. En su forma original, el instrumento consistía de 600 afirmaciones de cierto-falso que responde uno de los padres o la persona encargada del cuidado del niño. Las preguntas se parecen a las siguientes:

A mi hijo le resulta difícil quedarse dormido.
Mi hijo es melindroso para comer.
Mi hijo ha amenazado con matarse.
En ocasiones mi hijo insulta a otros adultos.
Nuestro matrimonio ha sido muy agitado.

La prueba también está disponible en tres versiones más cortas (131, 280 y 420 reactivos). Las cuatro versiones de la prueba difieren en cuanto al número de escalas que pueden calificarse. La versión completa con 600 reactivos produce cuatro escalas factoriales, cuatro escalas de validez, 12 escalas clínicas y 17 escalas experimentales. Las versiones de 280 y 420 reactivos no proporcionan calificaciones para las escalas experimentales, mientras que la versión de 131 reactivos, útil principalmente con propósitos de detección, produce sólo cuatro puntuaciones factoriales y un sólo índice de validez. En este caso se limitará el análisis a la versión con 420 reactivos.

Las escalas del PIC se detallan en el cuadro 14–8. Como ocurre con el MMPI, las puntuaciones naturales por escala se convierten en puntuaciones *T* con media de 50 y desviación estándar de 10. Las puntuaciones *T* elevadas aumentan la probabilidad de psicopatología o déficit. Las normas para niños de 6 a 16 años se basaron en una muestra de 2 390 niños y adolescentes que se obtuvo principalmente del sistema escolar en Minneapolis. Estos individuos provenían de antecedentes de clase social muy variados y proporcionaron una base razonable para la generalización de los resultados de la prueba. Las normas para niños preescolares, entre 3 y 5 años de edad, son mucho más limitadas (102 varones y 90 niñas) y la prueba debería utilizarse con precaución en este rango de edades.

Cuadro 14–8. Escalas del Inventario de personalidad para niños

Escalas factoriales
Indisciplinado/autocontrol deficiente
Incompetencia social
Internalización/Síntomas somáticos
Desarrollo cognitivo

Escalas de validez y detección
Mentira
Frecuencia
Defensividad
Adaptación

Escalas clínicas
Logro
Detección intelectual
Desarrollo
Preocupación somática
Depresión
Relaciones familiares
Delincuencia
Retraimiento
Ansiedad
Psicosis
Hiperactividad
Habilidades sociales

Con la posible excepción de las escalas de validez, los nombres de las escalas se explican por sí solos. Las escalas de validez incluyen Mentira, una escala de 15 reactivos diseñada para identificar a padres/encargados que son inocentemente defensivos acerca del niño; Frecuencia, una escala de 42 reactivos diseñada para identificar la exageración deliberada acerca de los síntomas o las respuestas aleatorias; Defensividad, una escala de 23 reactivos que se emplea para identificar defensividad acerca de

los problemas del niño; y Adaptación, una escala de 42 reactivos que proporciona un índice general de adaptación psicológica deficiente.

La mayor parte de las escalas del PIC se construyeron por medio de una variación estadísticamente compleja del enfoque de codificación de criterio empírico (Darlington, 1964; Darlington y Bishop, 1966). Este procedimiento de iteración añade los posibles reactivos de la escala según su correlación con el criterio pertinente (p. ej., delincuencia, psicosis, disfunción cerebral) y también considera las correlaciones entre reactivos para elegir la combinación de estos que produzca la correlación más alta entre la calificación de la escala y el criterio. Ocho de las escalas de perfil se construyeron a través de un método puramente racional basado en la nominación de reactivos de parte de expertos. Éstas fueron: Mentira, Desarrollo, Preocupación somática, Depresión, Relaciones familiares, Retraimiento, Ansiedad y Habilidades sociales.

La confiabilidad de las puntuaciones escalares del PIC es excepcionalmente buena, con coeficientes *test-retest* cuyo promedio se encuentra en la parte alta del rango de .80 para tres muestras diferentes (Wirt, Lachar, Klinedinst y Seat, 1984). Una monografía independiente de Lachar y Gdowski (1979) presenta los resultados confirmatorios de un amplio estudio de validez que se utilizó para establecer interpretaciones descriptivas de las calificaciones elevadas en las escalas del PIC. El estudio correlacionó los resultados de la prueba de 431 clientes de orientación infantil con los datos de criterio que proporcionaron padres, maestros y residentes psiquiátricos. Con base en estas conclusiones, se confirmaron varias interpretaciones descriptivas para las calificaciones elevadas en el PIC. El estudio de Lachar y Gdowski (1979) también sirvió de ímpetu para el desarrollo de un informe narrativo por computadora, disponible por correo si se solicita a los editores del PIC. Se encuentra al alcance una versión del programa para microcomputadora. Éste aplica, califica e interpreta el PIC. Segundos después de que se termina la prueba, el programa imprime un amplio resumen de los datos y un informe narrativo extenso.

Aunque el PIC tiene antecedentes que datan del decenio de 1950 (Wirt y Broen, 1958), el instrumento no atrajo la atención de los investigadores con orientación empírica sino hasta hace muy poco. En consecuencia, la investigación sobre la validez del PIC en realidad se encuentra en su infancia. No obstante, las conclusiones contemporáneas son bastante alentadoras. Por ejemplo, Keenan y Lachar (1988) probaron la eficacia del PIC en la identificación de niños preescolares que necesitan servicios de educación especial. Al utilizar una regla de decisión muy sencilla (dos o más escalas elevadas por arriba de $T = 70$), se clasificó de manera correcta a 91% de una muestra combinada, clínica y no clínica, de 120 niños. El PIC también es muy eficaz para identificar a niños de primaria con necesidades educativas especiales (Kline, Lachar y Boersma, 1993; Kline, Lachar, Gruber y Boersma, 1994). Lachar, Kline y Gdowski (1987) investigaron la precisión de los resultados del PIC en función de la psicopatología materna. Contrario a las expectativas de sentido común, las madres con perfiles clínicamente elevados en el MMPI tuvieron la misma precisión en la descripción de sus hijos que aquéllas con perfiles normales en el MMPI. Un problema del PIC es que el enfoque de clave de perfil —tan exitoso con el MMPI— ha resultado estéril en su caso (Kline, Lachar y Gdowski, 1992). Sin embargo, el PIC continúa siendo un instrumento útil cuando se interpreta escala por escala. Para revisiones adicionales véanse, Kline y Lachar (1992) y Rothermel y Lovell (1985).

RESUMEN

1. Los inventarios de autoinforme basados en la teoría dependen para su desarrollo de teorías explícitas de la personalidad. Un buen ejemplo de un inventario guiado por la teo-

ría es el *Edwards Personal Preference Shedule* (EPPS), un instrumento de 210 reactivos de elección forzosa que intenta medir las necesidades manifiestas de Murray a través de un autoinforme.

2. La *Personality Research Form* de Jackson (PRF) también se basó en el sistema de necesidades de Murray. Las 20 escalas de personalidad en la PRF no tienen superposición de reactivos y muestran una consistencia interna excepcional (mediana de .92). La validez de la PRF se fundamenta en análisis factoriales confirmatorios y en correlaciones apropiadas con escalas similares en otros instrumentos.
3. El Inventario Tipológico de Myers-Briggs (MBTI) es un inventario de autoinforme de elección forzosa que se basa de manera aproximada en la teoría de tipos de personalidad de Carl Jung. En el MBTI se califican cuatro dimensiones: Extraversión-Introversión, Sensación-Intuición, Pensamiento-Sentimiento y Juicio-Percepción, lo cual produce 16 tipos diferentes: como el ENSP.
4. El *Jenkins Activity Survey* (JAS) es un cuestionario de opción múltiple con 52 reactivos, diseñado para identificar el patrón conductual Tipo A, con propensión a la cardiopatía. Las tres subescalas incluyen: Velocidad e impaciencia, Participación laboral y Dinamismo y Competitividad. El JAS tiene varias limitaciones (p. ej., normas poco representativas, complejidades en la calificación) y, por tanto, es más adecuado para investigación.
5. Una prueba corta y sencilla que ha recibido altas calificaciones debido a su mérito técnico es el Inventario de ansiedad: rasgo-estado (IDARE). Los 40 reactivos de esta prueba se califican cada uno en una escala de intensidad de cuatro puntos. El IDARE mide ansiedad estado, o sentimientos transitorios de temor o preocupación; y ansiedad rasgo, la tendencia relativamente estable a responder de manera ansiosa ante situaciones estresantes.
6. El Cuestionario de 16 factores de la personalidad de Cattell (16FP) es típico de los instrumentos derivados del análisis factorial. Las cinco formas del 16FP (para grupos de diferentes edades) engloban todas un formato de elección forzosa. Los dieciséis atributos de personalidad examinados (y cuatro dimensiones de orden superior) se han confirmado de manera repetida a través del análisis factorial.
7. El Cuestionario de personalidad de Eysenck (EPQ) propone tres dimensiones principales de la personalidad, derivadas de manera analítico-factorial; Psicoticismo, Extraversión y Neuroticismo. Las confiabilidades de escala son bastante fuertes y la validez de constructo del instrumento ha recibido apoyo de docenas de estudios.
8. Las *Conrey Personality Scales* (CPS) constituyen un instrumento corto de autoinforme adecuado para estudiantes universitarios. Las ocho escalas CPS consisten de 20 reactivos cada una y no tienen superposición. Las escalas muestran excelente consistencia interna. Las calificaciones extremas son adecuadas en especial para predecir trastornos psicológicos.
9. El Inventario NEO de personalidad-Revisado (NEO PI-R) se basa en el modelo de cinco factores de la personalidad que se describe anteriormente. Los cinco constructos que mide la prueba son Neuroticismo, Extraversión, Apertura a la experiencia, Conformidad y Escrupulosidad. El NEO PI-R está disponible en dos formas paralelas que consisten de 240 reactivos calificados en una dimensión de cinco puntos.
10. El MMPI-2 consiste de 567 preguntas de cierto-falso. En la prueba se califican cuatro escalas de validez (?, L, F y K) que evalúan, respectivamente, las preguntas no respondidas, defensividad inocente, respuestas desviadas y defensividad sutil. Las 10 escalas clínicas son Hipocondriasis, Depresión, Histeria, Desviación psicopática, Masculinidad-feminidad, Paranoia,

Psicastenia, Esquizofrenia, Hipomanía e Introversión social.

11. El Inventario psicológico de California (CPI) es un instrumento parecido al MMPI que se diseñó para medir las dimensiones de la personalidad normal. Tres escalas miden las actitudes al responder prueba (p. ej., las tendencias a "fingirse bueno" y "fingirse malo"). Las 17 escalas clínicas se basan en conceptos "populares" de la personalidad que reconocen fácilmente los legos.
12. El Inventario multiaxial clínico de Millon, que ahora se encuentra en su tercera versión (MCMI-III) es una prueba corta (175 reactivos de cierto-falso) diseñada como auxiliar para el diagnóstico psiquiátrico. Las 27 escalas clínicas se organizaron en tres categorías amplias relacionadas con el DSM-IV: patrones clínicos de personalidad, patología grave de la personalidad, síndromes clínicos y síndromes clínicos graves.
13. El *Personality Inventory for Children* (PIC) se diseñó para proporcionar descripciones clínicamente apropiadas de la conducta infantil y de las características familiares. La forma más común consiste de 420 reactivos relacionados con el niño y la familia que responde uno de los padres o la persona encargada del cuidado del niño. Cuatro escalas proporcionan datos de validez y detección. La mayor parte de las 12 escalas clínicas se derivaron del enfoque de codificación de criterio.

TÉRMINOS Y CONCEPTOS CLAVE

Serie de respuestas de deseabilidad social p. 618
Prueba ipsativa p. 618
Ansiedad estado p. 623
Ansiedad rasgo p. 623
Extraversión p. 627
Introversión p. 627

TEMA 14B
Evaluación conductual y enfoques relacionados

Fundamentos de la terapia conductual
Terapia y evaluación conductual
Evaluación de conducta no verbal
Resumen
Términos y conceptos clave

En este tema, el lector se enfrentará con una variedad de enfoques directos, innovadores y, en ocasiones, no tradicionales de la evaluación de la personalidad, que de manera colectiva se conocen como **evaluación conductual**. Ésta se concentra en la conducta más que en los rasgos subyacentes, causas hipotéticas o supuestas dimensiones de la personalidad. Los muchos métodos de evaluación conductual ofrecen una alternativa práctica a las pruebas proyectivas, inventarios de autoinforme y otras técnicas inmanejables dirigidas a la evaluación global de la personalidad.

Por lo común, la evaluación conductual está diseñada para satisfacer las necesidades de los terapeutas y de sus clientes de manera rápida y sin complicaciones. Pero este tipo de evaluación difiere de la tradicional en más aspectos que su sencillez. Las suposiciones básicas, aspectos prácticos y metas esenciales de los enfoques conductual y tradicional son tan diferentes entre sí como el día y la noche. Las estrategias tradicionales de evaluación tienden a ser complejas, indirectas, psicodinámicas y, con frecuencia, ajenas al tratamiento. En contraste, las estrategias conductuales tienden a ser sencillas,

directas, de análisis de la conducta y continuas con respecto al tratamiento.

Los terapeutas conductuales utilizan una amplia gama de modalidades para evaluar a sus clientes, pacientes e individuos. Los métodos de evaluación conductual incluyen, pero no se limitan a: observaciones conductuales, autoinformes, calificaciones de padres, personal, hermanos, jueces, maestros, terapeutas y enfermeras; evaluación fisiológica, bioquímica y biológica; entrevistas estructuradas y semiestructuradas, y pruebas análogas. En su *Dictionary of Behavioral Assessment Techniques*, Hersen y Bellack (1988) listan 286 pruebas conductuales que se emplean en problemas y trastornos muy diversos en niños, adolescentes, adultos y población geriátrica. De modo que para que el lector pueda apreciar la diversidad de técnicas disponibles, se proporciona un muestrario de estas pruebas en el cuadro 14–9.

La evaluación conductual es, con frecuencia, —pero no siempre— una parte integral de la

Cuadro 14–9. Muestrario de pruebas y técnicas de evaluación conductual

Escala de movimientos anormales involuntarios
Evaluación de actividades cotidianas-modulares
Cuestionario de cogniciones agorafóbicas
Escala de creencias alcohólicas
Cuestionario de actividad antidepresiva
Situaciones asertivas
Prueba de autoafirmaciones de asertividad
Prueba de equiparación automatizada de figuras familiares
Escala de clasificación del perfil conductual
Evaluación conductual del bruxismo
Prueba de diapositiva de evitación conductual
Medidas conductuales de depresión grave
Prueba conductual de agudeza visual
Cuestionario de comilonas
Nivel de alcohol en sangre
Reactividad a la presión sanguínea
Cuestionario de sensaciones corporales
Sistema de codificación de la resistencia del cliente
Clasificación de demencia clínica
Escala de exposición al combate
Lista de verificación de actividad compulsiva
Cuestionario de conducta conflictiva
Diario del dormir cotidiano
Inventario Derogatis de funcionamiento sexual
Inventario de tentaciones alimenticias de Dieter
Escala de actitudes disfuncionales
Medición de monóxido de carbono en el aire espirado
Sistema de codificación de la interacción familiar
Escala de situaciones conductuales de urgencia en incendios
Prueba de competencia mental del tribunal de Georgia–Revisada
Escala de obtención de metas
Prueba de evitación de altura
Inventario de creencias irracionales
Inventario Leyton de obsesiones
Cuestionario McGill del dolor
Prueba Michigan de detección de alcohol
Escala de autoafirmaciones de ansiedad en la ejecución musical
Cuestionario de crisis de angustia
Sistema de codificación de la interacción padres-adolescentes
Respuestas de volumen del pene
Prueba de síntomas posteriores a la violación
Escala de calificación de autocontrol
Evaluación de autoafirmaciones a través de listado de pensamiento
Escala de búsqueda de sensaciones–Forma VI
Escalas de experiencia sexual
Lista de verificación de problemas verbales con el cónyuge
Caminata estandarizada
Inventario de probabilidad subjetiva de consecuencias
Alimentos de prueba en la evaluación de bulimia nerviosa
Inventario de valoración del tratamiento
Entrevista estructurada para personalidad Tipo A
Escala de analogía visual

Fuente: Basado en la lista de Hersen, M. y Bellack, A. S. (eds.) (1988). *Dictionary of behavioral assessment techniques*. Nueva York: Pergamon.

terapia conductual, diseñada para cambiar la duración, frecuencia o intensidad de una conducta blanco bien definida. Por ejemplo, una meta de la terapia para una estudiante universitaria tímida podría ser que ésta iniciara de manera diaria un mínimo de cinco conversaciones que duren dos minutos o más. El terapeuta podría recomendar que la estudiante enfocara esta meta de manera aumentativa, comenzando con unos cuantos intercambios sociales breves antes de proceder a conversaciones más extensas con personas desconocidas. En este ejemplo, la evaluación conductual podría asumir la forma de un automonitoreo en el que la estudiante utilizara un cronómetro para tomar el tiempo y un diario para llevar un registro de las conversaciones.

Como se señaló, es frecuente que la evaluación conductual esté al servicio de la terapia conductual. En muchos casos, la naturaleza de la evaluación conductual está en función de los procedimientos y metas de este tipo de terapia. Por tal razón, el lector podrá apreciar mejor las herramientas de la evaluación de la conducta si se entrelaza este tema con un análisis de los métodos de terapia conductual.

FUNDAMENTOS DE LA TERAPIA CONDUCTUAL

La terapia conductual, también llamada modificación de conducta, es la aplicación de los métodos y conclusiones de la psicología experimental a la modificación de la conducta desadaptada (Plaud y Eifert, 1998). Las raíces de la terapia conductual pueden encontrarse en *Science and Human Behavior* (1953), libro fundamental de Skinner en el que se detalla la aplicación del condicionamiento operante a los problemas de la conducta humana. Este autor rehuyó toda referencia a los acontecimientos privados, no observables, como pensamientos y sentimientos y en vez de ello enfatizó la importancia de identificar las conductas observables y de alterar de manera metódica las consecuencias ambientales de dichas conductas.

La investigación de Wolpe (1958) sobre el tratamiento conductual sistemático de las fobias también tuvo influencia en el establecimiento de los métodos de terapia conductual. Los procedimientos clínicos de Wolpe se obtuvieron de su trabajo de laboratorio sobre el condicionamiento y contracondicionamiento del temor en los gatos. Como Skinner, Wolpe le restó énfasis a la importancia de los pensamientos y creencias. Este experimentador consideraba que el temor es un fenómeno aprendido que puede desaprenderse al seguir un estricto protocolo de exposición por grados al objeto o situación temidos.

De manera más reciente, Bandura (1977), Mahoney y Arnkoff (1978) y Meichenbaum (1977) reintrodujeron los factores cognitivos al siempre cambiante esquema conductual. Por ejemplo, Bandura (1977) demostró que las personas son absolutamente capaces de un aprendizaje con base cognitiva. En particular, este autor mostró que los individuos pueden aprender de la mera observación de las contingencias de respuesta experimentadas por modelos. Dado que este aprendizaje existe en ausencia de consecuencias personales, debe tener una mediación cognitiva. Como consecuencia de este cambio de paradigma, casi todos los terapeutas conductuales de la actualidad se ocupan —al menos en cierto grado— de los pensamientos y creencias de sus clientes. Este nuevo énfasis se refleja en una familia de procedimientos de tratamiento muy populares conocidos de manera colectiva como terapia cognitiva–conductual (McMullin, 1986).

TERAPIA Y EVALUACIÓN CONDUCTUAL

En la actualidad, las técnicas específicas de terapia conductual pueden clasificarse en cinco categorías superpuestas (Johnston, 1986): métodos basados en la exposición, procedimiento de manejo de contingencias, terapias cognitivas-conductuales, procedimientos de autocontrol, y entrenamiento en asertividad y habilidades

sociales. La evaluación conductual se utiliza en todos estos enfoques, como se reseña en las siguientes secciones. No obstante, existen relativamente pocas herramientas con base conductual para la valoración de habilidades sociales, de modo que esta categoría no se analiza. Los lectores que deseen una cobertura limitada de instrumentos para la valoración conductual del entrenamiento en habilidades sociales (incluyendo asertividad) deberían consultar a Meier y Hope (1998).

MÉTODOS BASADOS EN LA EXPOSICIÓN

Los métodos de terapia conductual basados en la exposición son muy adecuados para el tratamiento de fobias que incluyen temores intensos e irracionales (p. ej., a las arañas, sangre, hablar ante un público). Un enfoque para la evitación fóbica es la exposición sistemática del cliente a la situación u objeto temidos. Wolpe (1973) favoreció la exposición gradual con un mínimo de ansiedad dentro de un procedimiento conocido como desensibilización sistemática. En este enfoque terapéutico, el cliente aprende primero a relajarse totalmente y después procede desde la exposición imaginaria hasta la actual o en vivo al estímulo temido. Otro método basado en la exposición es la inundación o implosión, en la que el cliente se sumerge de manera inmediata y total dentro de la situación inductora de ansiedad.

El terapeuta necesita cierto tipo de evaluación conductual para estimar el progreso continuo de un cliente que atraviesa por un tratamiento de las fobias basado en la exposición. En el enfoque de evaluación más sencillo posible, el cual se conoce como **Behavioral avoidance Test**, BAT (Prueba de Evitación Conductual; BAT), el terapeuta mide la cantidad de tiempo que el cliente puede tolerar al estímulo que le induce ansiedad. A continuación se presenta un clásico ejemplo de una BAT utilizada para evaluar a pacientes con agorafobia, un temor incapacitante a los espacios abiertos que con frecuencia se acompaña de ataques de pánico:

> La *Behavioral avoidance Test* (BAT) estandarizada se llevó a cabo una semana después del ingreso. Se habían retirado todos los medicamentos ansiolíticos, antidepresores y otros fármacos psicotrópicos al menos cuatro días antes de la prueba. El primer autor fue quien aplicó la prueba sin conocer los diagnósticos de los pacientes (y) sin haber participado en el tratamiento. Se pidió a los pacientes que caminaran solos tan lejos como pudieran del hospital, por un camino con tránsito ligero cuya longitud era de dos kilómetros. La ruta estaba dividida en ocho intervalos de extensiones iguales y los pacientes calificaron su ansiedad en una escala de 0 a 10 al final de cada intervalo. Los intervalos no concluidos recibieron una calificación de 10. Se calculó una calificación de evitación de la ansiedad al sumar las calificaciones de ansiedad de todos los intervalos (Hoffart, Friis, Strand y Olsen, 1994).

Los investigadores descubrieron que la calificación de evitación de la ansiedad en la técnica BAT se relacionaba en gran medida con los autoinformes de pensamientos catastróficos (p. ej., asfixiarse hasta morir, tener un infarto, actuar de manera absurda, quedar desamparado). Esta conclusión ilustra que, con frecuencia, los enfoques de evaluación conductual abarcan también un componente cognitivo. También debe señalarse la relación directa entre la meta de la terapia y la prueba de evitación conductual. En la agorafobia, la meta principal del tratamiento consiste en reducir la ansiedad de los pacientes al caminar por sí solos en espacios abiertos —que es exactamente lo que mide la BAT.

El enfoque BAT se basa en la suposición razonable de que el temor del cliente es el principal determinante de la conducta en la situación de prueba. Por desgracia, las características de la demanda de conducta deseable pueden ejercer una fuerte influencia sobre la conducta del cliente. La tolerancia de éste hacia el estímulo inductor de ansiedad tendrá cierta relación con el temor experimentado, pero también tienen mucho que ver con el contexto de la situación de evaluación (McGlynn y Rose, 1998). Puede ser que no se generalicen los resultados de las evaluaciones BAT y el terapeuta debe tener cautela de no concluir el tratamiento demasiado pronto.

El **Fear Survey Schedule** (Inventario de Estudio del Temor) es otro tipo de evaluación conductual, útil para la identificación y cuanti-

ficación de los temores. Los *Fear Survey Schedules* son instrumentos con validez de Facie que requieren que los individuos que los responden indiquen en una escala, por lo común del tipo Likert de 5 a 7 puntos, la presencia e intensidad de sus temores en relación con diversos estímulos. Se han publicado docenas de estos instrumentos, incluyendo versiones de Wolpe (1973), Ollendick (1983) y Cautela (1977). Tasto, Hickson y Rubin (1971) emplearon el análisis factorial para desarrollar un inventario de 40 reactivos que produce un perfil de calificaciones de temor en cinco categorías. En el cuadro 14–10 se muestra un inventario genérico de estudio del temor. Los *Fear Survey Schedules* se utilizan con frecuencia en proyectos de investigación para estudiar grandes muestras de personas en busca de aquellas que comparten un temor. Otro uso de estos inventarios consiste en registrar los cambios en temores, incluyendo aquellos que se han hecho el blanco de la intervención clínica.

Klieger y Franklin (1993) han expuesto varias advertencias acerca del uso de los *Fear Survey Schedules* en la investigación clínica. Estos autores señalan que los datos de confiabilidad para los estudios del temor son casi inexistentes. Un problema más grave tiene que ver con la validez de estos instrumentos. Al utilizar el *Fear Survey Schedule*-III (FSS-III) de Wolpe y Lang (1977), un inventario sumamente respetado y de amplio uso, Klieger y Franklin (1993) no encontraron relación entre los temores presentados en el FSS-III y medidas de BAT de los mismos temores. Por ejemplo, los individuos que informaban un fuerte temor a la sangre en el FSS-III estaban en la misma probabilidad de acercarse a una toalla blanca con sangre y tocarla que aquellos que no informaron temor a la sangre. Resultados similares se encontraron en personas que temían a serpientes, arañas y fuego. Los investigadores concluyeron que el FSS-III e instrumentos similares son una pobre elección para identificar a grupos experimentales y constituyen una base limitada para medir el resultado de las intervenciones terapéuticas. Su fracaso esencial parece residir en que estos inventarios poseen una validez tan "obvia", que pocos investigadores se han preocupado en evaluar las características psicomé-

Cuadro 14–10. Ejemplos de un Inventario de estudio del temor

Por favor marque en la columna que describa mejor su respuesta actual a estas situaciones u objetos.

	Grado al que usted se sentiría perturbado				
	De ninguna manera	Sólo un poco	Cantidad moderada	Mucho	Extremadamente alterado
Estar en un lugar desconocido					
Hablar en público					
Entrar a una fiesta					
Poner una inyección					
Que las personas me miren trabajar					
Grandes espacios abiertos					
Estar gordo					
Araña en la pared					
Gato en la habitación					
Regaño del jefe					

Nota: la mayoría de los inventarios de estudio del temor incluyen varias docenas de reactivos.

tricas tradicionales de confiabilidad y validez. Los *Fear Survey Schedules* deben utilizarse con cuidado.

PROCEDIMIENTOS DE MANEJO DE CONTINGENCIAS

El **manejo de contingencias** se basa en la suposición de que toda conducta —incluyendo la conducta trastornada o desadaptada— se conserva en función de sus consecuencias. El enfoque de manejo de contingencias para el cambio conductual procede en dos pasos. En el primero, el terapeuta intenta identificar las consecuencias que refuerzan de manera positiva una conducta indeseable. Por ejemplo, en el caso de un ama de casa con agorafobia, el terapeuta podría determinar que ella recibe la atención de su familia primordialmente cuando sufre pánico y está encerrada en casa. En esta situación, la conducta desadaptada de pánico y encierro se mantiene —al menos en parte— debido a la inoportuna amabilidad de los miembros de la familia. El paso dos consiste en cambiar las contingencias asociadas con la conducta indeseable. El terapeuta podría recomendar que los miembros de la familia no le presten atención al ama de casa durante sus episodios de pánico y que refuercen sus pequeños pasos hacia la independencia. Por ejemplo, si la mujer camina al buzón de correos y de regreso, los miembros de la familia deberían demostrarle su atención.

Una forma de manejo de contingencias que se utiliza con amplitud en instituciones es la economía de fichas. Este enfoque se adapta de manera particulamente adecuada a clientes con repertorios conductuales limitados. En una **economía de fichas**, muchos tipos diferentes de conducta prosocial se recompensan a través de fichas que después pueden intercambiarse por recompensas materiales o privilegios (Kazdin, 1988).

La evaluación conductual en una economía de fichas asume principalmente la forma de observación conductual directa (Foster, Bell-Dolan y Burge, 1988). La aplicación de un sistema de observación conductual directa es mucho más demandante de lo que parece. El investigador o terapeuta debe identificar las conductas blanco, definirlas de manera precisa, producir un sistema para el registro de datos, calibrar las recompensas apropiadas y entrenar a los miembros del personal (Paul, 1986). Si el programa funciona, la frecuencia observada de conductas prosociales determinadas aumentará de modo gradual a lo largo de un periodo de meses o semanas.

Las economías de fichas fueron muy populares en el decenio de 1970, cuando se les empleaba incluso en sistemas escolares completos para facilitar la mejoría en habilidades sociales de estudiantes de primaria que sufrían desventajas (Bushell, 1978). Las economías de fichas siguen siendo populares, pero el entusiasmo inicial acrítico se ha visto reemplazado por una perspectiva menos jovial. Kazdin (1988) analiza las limitaciones de las economías de fichas, las cuales incluyen: 1) generalización deficiente de las conductas blanco más allá del programa de tratamiento, 2) dificultades para entrenar al personal para que ponga en marcha de manera adecuada las economías de fichas y 3) resistencia del cliente a la participación. Además, la reactividad puede ser un problema en la observación conductual directa: es posible que los clientes no se comporten de manera natural cuando saben que se les observa.

TERAPIAS COGNITIVAS-CONDUCTUALES

El factor común a todos los tipos de terapia cognitiva-conductual es un énfasis en el cambio de la estructura de creencias del cliente. Las tres variedades más conocidas de **terapia cognitiva-conductual** son la terapia racional emotiva (TER) de Ellis (1962), el entrenamiento de autoinstrucción de Meichenbaum (1976) y la terapia cognitiva de Beck (1976). Ellis postula que la mayor parte de la conducta trastornada es el producto de creencias irracionales, como la muy difundida creencia de que uno debe tener el amor y la aprobación de todas las personas significativas en todo momento. Ellis intenta alterar dichas creencias irracionales centrales

principalmente a través de los argumentos lógicos y la exhortación contundente. La técnica de autoinstrucción de Meichenbaum consiste en enseñarle al cliente a utilizar autoafirmaciones de enfrentamiento para combatir las situaciones estresantes. Por ejemplo, a un estudiante universitario que sufre de una intensa ansiedad cuando presenta pruebas se le podría enseñar a utilizar la siguiente plática interna durante los exámenes: "Esta vez tienes una estrategia... Respira profundamente y relájate... Tan sólo responde a una pregunta a la vez...". La terapia cognitiva de Beck se concentra de manera primordial en el papel de las distorsiones cognitivas en el mantenimiento de la depresión y otros trastornos emocionales. Beck (1983) considera que la depresión es, de manera fundamental, un trastorno cognitivo que se caracteriza por la tríada cognitiva negativa: una perspectiva pesimista del mundo, un autoconcepto pesimista y un punto de vista pesimista acerca del futuro. En la terapia, Beck utiliza una forma cordial de reestructuración cognitiva para ayudar al cliente a percibir sus problemas en términos alternativos, que tengan una posible solución.

Los terapeutas cognitivo-conductuales no necesitan utilizar herramientas formales de evaluación en su práctica clínica. Por lo común, estos terapeutas vigilan la estructura de creencias de sus clientes de manera informal de una sesión a otra. Los pensamientos irracionales y distorsionados se ponen en duda a medida que surgen en la terapia. Al final, el autoinforme de mejoría de parte del cliente puede constituir el principal índice de éxito terapéutico. No obstante, se encuentran disponibles varias medidas directas de la distorsión cognitiva. En el cuadro 14–11 se resumen unos cuantos instrumentos destacados. Pueden encontrarse otros ejemplos en Clark (1988). Estos instrumentos son principalmente cuestionarios de investigación adecuados para el examen de las diferencias individuales, pero no se han validado de manera suficiente para la evaluación individual. Clark (1988) critica a los creadores de los cuestionarios de distorsión cognitiva de publicar prematuramente sus instrumentos. En particular, este autor señala la falta de investigación sobre validez concurrente y de discriminación en la mayor parte de las medidas de autoafirmaciones. Otro problema es que los cuestionarios existentes se diseñaron para validar constructos en investigación y, en consecuencia, no funciona bien en la práctica clínica.

Una medida excepcional y bien validada que no se encuentra incluida en el cuadro 14–11 es el *Beck Depression Inventory*; BDI (Inventario de Depresión Beck). El BDI es un cuestionario corto y sencillo de autoinforme que se enfoca, en parte, en las distorsiones cognitivas que subyacen a la depresión (Beck y Steer, 1987; Beck, Ward, Mendelsohn, Mock y Erbaugh, 1961). Una razón para su popularidad es que la mayoría de los individuos pueden contestar los 21 reactivos del BDI en 10 minutos o menos. La prueba se ha utilizado ampliamente y se han publicado más de 1 900 artículos que utilizan el BDI (Conoley, 1992). En 1996 se publicó una segunda edición (Beck, Steer y Brown, 1996). En el BDI-II, se revisaron varios reactivos a fin de lograr que el inventario se conformara de manera más cercana a los criterios diagnósticos prevalecientes para la depresión. Los 21 reactivos se presentan de la siguiente forma:

Marque la afirmación que siente más cierta acerca de usted dentro de este grupo:

0 Soy optimista acerca del futuro.
1 Siento un leve desaliento acerca del futuro.
2 Siento que el futuro tiene poco que ofrecerme.
3 Siento que no hay ninguna esperanza en el futuro.

Trece reactivos abarcan componentes cognitivos y afectivos de la depresión como pesimismo, culpa, llanto, indecisión y autoacusaciones; ocho evalúan variables somáticas y de desempeño como problemas de sueño, imagen corporal, dificultades laborales y pérdida de interés en el sexo. La persona examinada recibe una calificación de 0 a 3 para cada reactivo; la puntuación natural total es la suma de las afirmaciones refrendadas en los 21 reactivos; la calificación máxima posible es 63.

Cuadro 14–11. Medidas de cuestionario de la distorsión cognitiva

Anxious Self-Statements Questionnaire **(ASSQ) (Cuestionario de Autoafirmaciones de Ansiedad)**
Kendall y Hollon (1989)

La persona examinada califica con qué frecuencia ocurrieron pensamientos específicos de ansiedad a lo largo de la última semana. Los reactivos tienen la siguiente forma:

Ya no lo puedo soportar
¿Ahora qué me va a suceder?
No voy a lograrlo

El ASSQ es un instrumento psicométricamente sólido y puede utilizarse para evaluar los cambios en la frecuencia de la plática ansiosa interna.

Automatic Thoughts Questionnaire **(ATQ) (Cuestionario de Pensamientos Automáticos)**
Hollon y Kendall (1980); Kazdin (1990)

El ATQ es una medida de la frecuencia de las cogniciones relacionadas con la depresión que evalúa la desadaptación personal, negatividad en autoconcepto y expectativas, baja autoestima y desamparo/darse por vencido. El ATQ tiene 30 reactivos y se correlaciona muy bien con la Escala de depresión del MMPI y el *Beck Depression Inventory* (Ross, Gottfredson. Christensen y Weaver, 1986).

Cognitive Errors Questionnaire **(CEQ) (Cuestionario de Errores Cognitivos)**
Lefebvre (1981)

El CEQ evalúa el grado de pensamiento desadaptado en situaciones generales y también en aquéllas relacionadas con lumbalgia crónica. A una serie de viñetas separadas relacionadas con la lumbalgia crónica y escenas generales le siguen cogniciones disfóricas ilógicas. La persona que responde al cuestionario indica en una escala de cinco puntos qué tan similar es la cognición al pensamiento que pudiera tener en la misma situación. Por ejemplo: "Usted acaba de pasar tres horas limpiando el sótano de su casa. Sin embargo, su cónyuge no dice nada al respecto. Usted piensa para sí mismo(a). Él (ella) debe creer que hice un mal trabajo". Smith, Follick, Ahern y Adams (1986) encontraron que la generalización excesiva era el error cognitivo específico del CEQ que se correlacionaba de manera más consistente con la discapacidad por lumbalgia crónica.

Attribution Styles Questionnaire **(ASQ) (Cuestionario de Estilos de Atribución)**
Seligman, Abramson, Semmel y Von Baeyer (1979)

El ASQ mide tres dimensiones de atribución relacionadas con el modelo de desamparo aprendido de Seligman para la depresión: dimensiones interna-externa, estable-inestable, y global-específica. Las personas depresivas atribuyen los malos resultados a causas internas, estables y globales, mientras que atribuyen los buenos resultados a causas externas e inestables. El cuestionario consiste de 12 situaciones hipotéticas, seis de las cuales describen buenos resultados y seis que describen malos resultados (p. ej., "Durante algún tiempo, usted ha estado buscando empleo sin tener éxito"). La persona califica cada viñeta en una escala de siete puntos según el grado de internalidad, estabilidad y globalidad.

Hopelessness Scale **(HS) (Escala de Desesperanza)**
Beck (1987), Dyce (1996)

La HS es una escala con 20 reactivos de cierto/falso que se diseñó para cuantificar la desesperanza, un componente de la tríada cognitiva negativa que se encuentra en las personas depresivas. (La tríada incluye las perspectivas negativas acerca de sí mismo, el mundo y el futuro.) La escala es sensible a cambios en el estado depresivo del paciente. En un estudio de validación, Beck, Riskind, Brown y Steer (1988) encontraron que las calificaciones en la HS tenían una relación insignificante con la ansiedad o la psicopatología en general cuando se separaba la influencia de depresión coexistente. Así, la HS parece medir un atributo específico de depresión más que la psicopatología general.

En un metaanálisis de los estudios de investigación con el BDI, la consistencia interna de la escala (coeficiente alfa) fue de .73 a .95, con una media de .86 en nueve poblaciones psiquiátricas (Beck, Steer y Garbin, 1988). El BDI-II posee excelente consistencia interna con un coeficiente alfa de .92 (Beck, Steer y Brown, 1996). La confiabilidad *test-retest* del BDI es modesta, con un rango de .60 a .83 en muestras no psiquiátricas y de .48 a .86 en muestras psiquiátricas. Sin embargo, la metodología *test-retest* no es muy adecuada para fenómenos como la depresión que son inestables por naturaleza. La depresión subjetiva fluctúa de manera notable de una semana a la siguiente, día a día e incluso de una hora a otra. Un valor mediocre de confiabilidad *test-retest* podría significar un cambio válido en el constructo que se mide más que un error indeseable de medición.

Se encuentra disponible una variedad de resultados normativos con datos del BDI para muestras de pacientes con depresión mayor, distimia, alcoholismo, adicción a la heroína y problemas mixtos. El manual también proporciona pautas para el grado de depresión con base en la calificación del BDI (0 a 9, normal; 10 a 19, leve a moderada; 20 a 29, moderada a grave; 30 en adelante, sumamente grave). Estas calificaciones se basan en las valoraciones clínicas de los pacientes.

El BDI se ha validado de manera extensa contra otras medidas y criterios independientes de la depresión. Por ejemplo, las correlaciones con calificaciones clínicas y escalas de depresión como la del MMPI se encuentran por lo común en el rango de .60 a .76 (Conoley, 1992). Las diferencias sexuales son mínimas, aunque existen diferencias leves en la expresión de la depresión entre varones y mujeres (Steer, Beck y Brown, 1989).

El único defecto del BDI es su transparencia. Los pacientes que desean ocultar su desesperación o exagerar su depresión pueden hacerlo con facilidad. No obstante, en el caso de pacientes motivados a reflejar de manera precisa su estado emocional, es probable que el BDI y el BDI-II sean invencibles como índices de la presencia y grado de la depresión (Stehouwer, 1987). Algunos profesionales le piden a los pacientes que respondan al BDI después de cada sesión terapéutica; lo utilizan en mucho como podrían emplear un termómetro.

PROCEDIMIENTOS DE AUTOMONITOREO

Una concepción errónea común acerca de la terapia conductual es que consiste de un terapeuta autoritario que le aplica recompensas y castigos poderosos a los clientes pasivos. Aunque este modelo estereotipado puede ser cierto con algunos clientes discapacitados que tienen repertorios conductuales limitados, en su mayor parte la terapia conductual la realizan profesionales humanitarios que le enseñan a sus clientes métodos de autocontrol. En todas las formas de terapia conductual es fundamental un énfasis en el **automonitoreo**. En éste, el cliente elige las metas y participa de manera activa en supervisar, esquematizar y registrar el progreso hacia la finalidad o finalidades de la terapia. Según este modelo, al terapeuta se le relega a un papel de asesor experto.

Los procedimientos de automonitoreo son especialmente útiles en el tratamiento de la depresión, un trastorno conductual preponderante que consiste de estado de ánimo triste, bajo nivel de actividad, sentimientos de minusvalía, problemas de concentración y síntomas físicos (pérdida de sueño, trastorno del apetito, menor interés en el sexo). Se han presentado varios programas de automonitoreo para la depresión (Lewinsohn y Talkington, 1979; Rehm, 1984; Rehm, Kornblith, O'Hara y cols., 1981). A fin de ilustrar el enfoque del automonitoreo para el control de la depresión, se resumirá un pequeño trozo del programa que impulsan Lewinsohn y sus colaboradores (Lewinsohn, Munoz, Youngren y Zeiss, 1986).

Lewinsohn observó que la depresión va de la mano con una notable reducción en la experiencia de acontecimientos agradables. Las personas deprimidas se alejan de participar en actividades agradables; el retraimiento conductual sólo contribuye a aumentar su depresión, incitando a que se produzca una continua espiral

hacia abajo. Por fortuna, es posible reemplazarla con una espiral hacia arriba. Para ayudar a revertir la espiral invertida de la depresión, Lewinsohn y sus colaboradores diseñaron el *Pleasant Events Schedule*, PES (Inventario de Acontecimientos Placenteros; MacPhillamy y Lewinsohn, 1982). El propósito del PES es doble. Primero, en la fase de evaluación de línea base, esta prueba se utiliza para automonitorear la frecuencia (F) y agradabilidad (A) de 320 acontecimientos comunes, cotidianos. Ejemplos de los tipos de acontecimientos que se listan en el PES incluyen:

Lectura de revistas.
Salir a caminar.
Estar con mascotas.
Tocar un instrumento musical.
Preparar comidas para caridad.
Escuchar el radio.
Leer poesía.
Asistir a servicios religiosos.
Mirar un evento deportivo.
Jugar baraja con un amigo.
Trabajar en mi empleo.

La frecuencia y agradabilidad de estos hechos cotidianos se califican de 0 a 2.* A partir de esto se calcula la tasa media de actividades placenteras al sumar las calificaciones F × A, es decir, tasa media = F × A/320. Los datos normativos para la media de F, media de A y media de F × A se presentan en Lewinsohn, Munoz, Youngren y Zeiss (1986) y sirven como base para planificar el tratamiento. Los participantes en el programa de Lewinsohn también registran su estado de ánimo diario de manera sencilla de 0 (peor) a 9 (mejor).

La segunda utilización del PES es automonitorear el progreso terapéutico. Con base en los resultados iniciales de la prueba, los clientes identifican aproximadamente 100 acontecimientos agradables en potencia y se esfuerzan por aumentar la frecuencia de éstos, registrando su estado de ánimo de manera cotidiana. En general, los clientes que aumentan la frecuencia de hechos placenteros muestran una mejoría en estado de ánimo y en otros síntomas depresivos.

El PES es una herramienta muy útil para los clínicos que desean poner en marcha el enfoque de automonitoreo para la evaluación y tratamiento de la depresión. MacPhillamy y Lewinsohn (1982) informan conclusiones favorables acerca de las cualidades técnicas del PES y analizan una variedad de subescalas racionales, factoriales y empíricas que no pueden revisarse aquí. El instrumento tiene una confiabilidad *test-retest* que va de adecuada a buena (correlaciones a un mes dentro del rango de .69 a .86), excelente validez concurrente con observadores entrenados y una prometedora validez de constructo. En general, las subescalas se comportan como podría predecirse con base en los constructos que pretenden medir —se refiere al lector a MacPhillamy y Lewinsohn (1982) para mayores detalles.

Los enfoques de automonitoreo para la evaluación conductual tienen aplicaciones en muchas subespecialidades de la psicología y de las profesiones de salud. Se describirá de manera breve otro instrumento de evaluación para ilustrar la diversidad de métodos disponibles. Schlundt y Bell (1993) desarrollaron el *Body Image Testing System* (Sistema de prueba de la imagen corporal; BITS) para utilizarse con microcomputadora, que tiene el propósito de evaluar la imagen corporal de personas con trastornos de la alimentación. El trastorno alimenticio conocido como anorexia nerviosa se caracteriza por la negativa a mantener un peso corporal normal, un intenso temor a subir de peso y un trastorno significativo de la percepción de la forma o tamaño del cuerpo. La ma-

* La Escala de frecuencia se calibra como sigue:

0 —Esto **no** ha sucedido en los pasados 30 días.
1 —Esto ha sucedido **unas cuantas veces** (1 a 6 veces) en los pasados 30 días.
2 —Esto ha sucedido **con frecuencia** (7 veces o más) en los pasados 30 días.

La Escala de agradabilidad se calibra como sigue:

0 —Esto **no** fue agradable.
1 —Esto fue **un poco** agradable.
2 —Esto fue **muy** agradable.

yoría de los pacientes con anorexia nerviosa son mujeres jóvenes —la edad promedio de inicio es de 17 años. Una persona con este trastorno percibe su cuerpo como más grueso de lo que es en realidad. Pero ¿cómo puede medirse y cuantificarse esta percepción equivocada? El programa BITS es un enfoque para el automonitoreo de las distorsiones en la imagen corporal. En pocas palabras, la persona utiliza un menú para cambiar una imagen corporal generada por computadora hasta que se asemeje a su percepción de sí misma.

Las instrucciones para el programa BITS especifican que la sujeto ha de ajustar las partes individuales del cuerpo hasta que la imagen se parezca a sus dimensiones corporales reales. En un segundo ensayo, la sujeto genera dimensiones corporales ideales. La discrepancia entre las dimensiones reales e ideales representa el grado de satisfacción de la sujeto con su propio cuerpo. Además, el programa BITS produce un índice de distorsión perceptual que se basa en un análisis de regresión de la altura y peso contra la imagen corporal autogenerada (real).

Al utilizar los datos de 94 mujeres, estudiantes de pregrado, a quienes se volvió a someter a prueba después de dos a cuatro semanas, se encontró que las medidas del BITS tenían una confiabilidad aceptable. Por ejemplo, las calificaciones de satisfacción general para las nueve partes del cuerpo mostraron una confiabilidad *test-retest* de .80. También están disponibles los datos normativos de 528 individuos, obtenidos principalmente de muestras de pregrado. Se encontró que la validez de constructo es muy prometedora, con base en tres fuentes de información: fuertes correlaciones entre las calificaciones del BITS y el tamaño corporal real; fuertes correlaciones entre las variables del BITS y otras medidas de la imagen corporal; y capacidad de la medida para predecir trastornos en la alimentación como hacer dietas, comer en exceso y comer para satisfacer necesidades emocionales.

El procedimiento BITS es un ejemplo excelente de una nueva generación de pruebas psicológicas que se han vuelto posibles por los recientes avances en la tecnología de cómputo. Las pruebas que le dan un uso creativo a los gráficos por computadora para presentar estímulos, respuestas o ambos, se volverán cada vez más comunes en el futuro. En el Tema 15A se analizan los enfoques adicionales para la evaluación psicológica asistida por computadora.

EVALUACIÓN DE LA CONDUCTA NO VERBAL

La **conducta no verbal** incluye las formas sutiles de comunicación humana contenidas en la mirada, ademanes, lenguaje corporal, tono de voz y expresión facial. Aunque la comunicación no verbal representa un papel crucial en la conducta humana, nuestro conocimiento de ella es muy incompleto. Esta ignorancia refleja, en parte, la fuerte orientación verbal de nuestra sociedad; se hace una equivalencia entre comunicación y empleo exitosos de las palabras. Pero existen otras razones más sutiles para la falta de conocimiento científico acerca de la comunicación no verbal:

> Los diferentes tipos de comunicación no verbal se encuentran tan incrustados en nuestras vidas cotidianas que utilizamos los mensajes no verbales sin tener conciencia de ellos. Por ejemplo, cuando nos formamos una opinión acerca de cómo es una persona, es probable que dicha opinión se base en parte en un análisis complejo de la información no verbal. Cuando concluimos que alguien a quien acabamos de conocer está enojado o celoso o ansioso por irse, es posible que hayamos llegado a esta conclusión tanto por escuchar el tono de voz de la persona y observar qué tan agitados son sus movimientos o por habernos formado una impresión de la calidez de su expresión facial como por interpretar lo que en realidad se dijo (Rosenthal, Hall, DiMatteo, Rogers y Archer, 1979).

En esta sección se investigará el estudio científico de la comunicación no verbal, con un enfoque sobre las herramientas y métodos que pueden utilizarse con propósitos de evaluación. Se comenzará con una revisión de los métodos que se utilizan para discernir la naturaleza de la personalidad subyacente a partir de signos

conductuales externos como la interacción visual, las señales paralingüísticas (p. ej., tono de voz) y la expresión facial. El capítulo cierra con un análisis del *Perfile of Nonverbal Sensitivity* (Perfil de sensibilidad no verbal; PONS), una fascinante prueba con película o vídeo para evaluar la sensibilidad a la comunicación no verbal de otras personas.

INTERACCIÓN VISUAL

Desde hace mucho tiempo se ha reconocido que la interacción visual es clave para otros aspectos de la personalidad, aunque sólo ha sido en los últimos decenios que ha aparecido la investigación sistemática sobre este tema. Nielsen (1962) llevó a cabo la primera investigación empírica acerca de la mirada en la conducta social. Su estudio fue puramente un análisis observacional de la confrontación social y la interacción visual. Poco después, otros investigadores comenzaron a realizar estudios programáticos en los que se emplearon diseños experimentales estandarizados, con el contacto visual como variable dependiente y, después, como variable independiente (resumidos en Argyle y Cook, 1976; Fehr y Exline, 1987; Exline y Fehr, 1982). En gran medida, esta investigación ha buscado determinar las implicaciones de diferencias individuales como el inicio, mantenimiento y comodidad con el contacto visual individual y mutuo. De estos estudios han surgido unas cuantas generalizaciones frágiles que se resumirán más adelante. No obstante, es más necesario mencionar el consenso básico acerca de la dificultad para obtener medidas significativas de la interacción visual. Permítasenos revisar de manera breve las conclusiones y después se reseñarán los obstáculos para la investigación útil sobre este importante tema.

Es mejor considerar a la interacción visual como una conducta social que actúa como una poderosa señal de una persona a otra. El significado de la señal se determina de muchas maneras complejas, de modo que la interpretación del patrón de mirada de otra persona es muy difícil. Sin embargo, sí han surgido varias conclusiones consistentes acerca de la investigación sobre el tema (Argyle y Cook, 1976; Fehr y Exline, 1987):

1. La mirada actúa con frecuencia como una señal de agrado, en especial por parte de personas dependientes.
2. La mirada mutua, en especial si es prolongada, puede significar un tipo especial de intimidad.
3. Las personas extravertidas y que tienen confianza en sí mismas establecen contacto visual con mayor frecuencia y por periodos más largos.
4. La aversión a la mirada aumenta cuando las personas están cercanas entre sí o discuten temas íntimos.
5. Las personas poco honradas reducen la interacción visual durante episodios de engaño.
6. Las mujeres fijan la mirada con más frecuencia, incluso en la lactancia.
7. La cantidad de interacción visual es alta durante la infancia, disminuye en la adolescencia y después vuelve a elevarse.
8. Las personas afiliativas fijan más la mirada en situaciones cooperativas, pero menos en situaciones competitivas.
9. Las personas con un estatus inferior fijan más la mirada que las personas con mayor estatus.
10. Existen notables diferencias culturales; las personas árabes y latinoamericanas tienen mayor interacción visual; ciertos grupos amerindios tienen menor interacción visual.
11. Las personas con autismo, esquizofrenia y depresión tienden a evitar mirar a otras personas.

De la lista anterior debería ser obvio que la interpretación de la interacción visual depende de numerosas variables y también está bajo la influencia de la interacción entre variables (p. ej., afiliación y competencia).

El profesional que desee interpretar las diferencias individuales en interacción visual primero debe poner en práctica un paradigma de

algún tipo para la recolección de datos. Exline y Fehr (1982) han llamado la atención hacia las dificultades implicadas en la evaluación de la mirada y de la mirada mutua. Estas dificultades incluyen la reactividad a las mediciones realizadas en ambientes de laboratorio y los problemas de confiabilidad de las mediciones. Son muchos los problemas prácticos para evaluar e interpretar la interacción visual. De hecho, es muy probable que los obstáculos impidan el desarrollo de pruebas estandarizadas para la medición de la mirada y de la mirada mutua. No obstante, el conocimiento acerca de los hallazgos en esta área puede utilizarse junto con otras señales no verbales en aplicaciones clínicas prácticas como la detección del engaño (Ekman y Friesen, 1975).

PARALINGÜÍSTICA

La **paralingüística** se refiere al tono de voz, tasa de verbalizaciones y otros aspectos no verbales del habla. Con frecuencia, las señales paralingüísticas son más poderosas que el mensaje hablado explícito, como cuando se dice: "No es lo que dijo, sino cómo lo dijo". Una máxima esencial de la investigación paralingüística es que se debe separar el contenido del discurso de sus matices afectivos.

Una importante área de la evaluación paralingüística tiene que ver con la emoción juzgada a partir del discurso con "filtración del contenido". En éste, se graba la voz de una persona (o se vuelve a grabar una cinta de audio) a través de un filtro de baja frecuencia para eliminar los sonidos de alta frecuencia y con ello hacer que las palabras se vuelvan irreconocibles. El discurso con filtración de contenido mantiene los aspectos paralingüísticos esenciales del habla (variación en tono y contorno; ritmo, volumen, tonalidad y ritmo), pero elimina la distracción potencial del contenido. Jueces entrenados califican entonces el discurso con filtración de contenido en cuanto a componentes afectivos como enojo o ansiedad.

Milmoe, Rosenthal, Blane, Chafetz y Wolf (1967) llevaron a cabo un estudio que ha servido de guía y que ilustra las aplicaciones prácticas del discurso con filtración de contenido. Estos investigadores grabaron las voces de médicos residentes que habían terminado su servicio en una clínica de alcoholismo. En particular, los investigadores obtuvieron los fragmentos de las respuestas de los médicos a la pregunta de entrevista, "¿Cuál ha sido su experiencia con alcohólicos?". En éstas se filtró el contenido y después un panel de jueces calificó la presencia de enojo, compasión, ansiedad y neutralidad* en una escala de 6 puntos (1 = ninguno, 6 = mucho). La confiabilidad intercalificadores para las dimensiones emocionales varió en gran medida y también se vio afectada por el sexo del juez. Por ejemplo, la concordancia entre los jueces varones y mujeres fue buena en cuanto a ansiedad ($r = .84$), pero no existente para neutralidad ($r = -.07$).

Los investigadores también obtuvieron información ulterior sobre el éxito de los médicos para lograr que los pacientes alcohólicos buscaran tratamiento. Los datos paralingüísticos (calificaciones de enojo, compasión, ansiedad y neutralidad) se correlacionaron entonces con los datos objetivos de eficacia en la canalización. Las calificaciones de los jueces para el enojo en el discurso con filtración de contenido se correlacionaron en gran medida con la eficacia para canalizar a los pacientes alcohólicos ($r = -.67$, $p = .06$). Otras variables no tuvieron relación con la eficacia. En síntesis, los médicos con altas calificaciones paralingüísticas de enojo fueron poco efectivos para convencer a los pacientes alcohólicos de que buscaran tratamiento. El mismo equipo de investigación ha mostrado que las calificaciones de la voz materna (con filtración de contenido) se correlacionan con aspectos de la conducta del recién nacido como irritabilidad, inseguridad y atención (Milmoe, Novey, Kagan y Rosenthal, 1974).

* Otras dos dimensiones se calificaron a partir del **contenido** de las cintas: complejidad y comprensión psicológica. En este caso no se analizan estas conclusiones.

EXPRESIÓN FACIAL

La comunicación no verbal se media con frecuencia por la expresión facial. De hecho, el aparato de expresión facial se dedica en gran medida a la manifestación facial de emoción. Además, el rostro ayuda a regular varias funciones no emocionales como la producción del discurso, la alimentación, respiración, olfato y protección de los ojos.

El papel crucial del rostro en la comunicación a través de la presentación de emociones ha impulsado a los investigadores a desarrollar métodos objetivos para observar y cuantificar la acción facial. Han surgido dos métodos importantes para este propósito: 1) medición de las acciones faciales visibles utilizando sistemas de codificación facial y 2) medición de descargas eléctricas producidas por la contracción de los músculos faciales. Ekman (1982) y Fridlund, Ekman y Oster (1987) revisaron en detalle ambos desarrollos. Aquí se destaca el sistema de codificación facial más conocido.

El *Facial Action Codying System*, FACS (Sistema de codificación de la acción facial; Ekman y Friesen, 1978) se desarrolló como una herramienta de evaluación con fines generales adecuada para una amplia gama de propósitos de investigación y evaluación. Con base en un elaborado análisis electrofisiológico del papel preciso de cada músculo del rostro en la expresión facial visible, Ekman y Friesen obtuvieron 44 unidades de acción (UA) que pueden, solas o en combinación, explicar todos los movimientos faciales visibles. Todas las UA reciben una calificación en una escala de intensidad de cinco puntos. Por ejemplo, la UA 1 tiene que ver con levantar la ceja, lo cual se controla a través de un músculo grande en el área de la frente. Ekman y Friesen (1978) proporcionan instrucciones detalladas para calificar cada UA a partir de grabaciones de vídeo. El FACS también permite la codificación del tiempo de inicio, cúspide y terminación de cada UA.

El FACS es difícil de aprender y de utilizar debido a que requiere ver de manera repetida y en cámara lenta las acciones faciales. No obstante, una vez que se domina el sistema, jueces competentes producen calificaciones sumamente confiables de las acciones faciales. Lo que es más, las calificaciones predicen estados emocionales con una precisión considerable. En un importante estudio de validez, Ekman, Friesen y Ancoli (1980) registraron la acción facial de personas que observaban películas tanto agradables como desagradables. El FACS predijo de manera precisa los informes retrospectivos de los individuos acerca de su experiencia emocional (felicidad, sentimientos negativos, repugnancia). Al menos en algunos contextos, las acciones faciales particulares sí indican emociones particulares. Lo que necesita establecerse en la investigación futura es la generalidad del vínculo entre las acciones faciales codificadas del FACS y la experiencia de emociones pronosticadas.

PERFIL DE SENSIBILIDAD NO VERBAL

Robert Rosenthal y sus colaboradores desarrollaron *Profile of Nonverbal Sensitivity*, PONS (Perfil de Sensibilidad no Verbal), en el decenio de 1970 para estudiar la capacidad para comprender las señales no verbales trasmitidas por expresiones faciales, movimientos corporales y tono de voz (Rosenthal, Hall, DiMatteo, Rogers y Archer, 1979). Un impulso para la invención de este nuevo instrumento de evaluación fue el descubrimiento anterior de Rosenthal acerca del efecto de la expectativa del experimentador (Rosenthal, 1967). La investigación sobre el efecto de la expectativa del experimentador mostró que el investigador puede influir sin intención sobre los individuos y cambiar con ello su conducta en dirección de las expectativas del experimentador, con lo cual se crea una profecía autocumplida en el laboratorio de investigación y en el salón de clase. Rosenthal razonó que el efecto de la expectativa del experimentador debe facilitarse a través de la comunicación involuntaria de información no verbal del experimentador al individuo. El PONS se desarrolló, en parte, para estudiar la manera en que podrían decodificarse estas comunicaciones no verbales.

Los estímulos del PONS se aplican por medio de película o videocinta y consisten de 220 segmentos de dos segundos acerca de la conducta no verbal de una mujer.* Para cada reactivo estímulo, la persona examinada debe elegir qué situación afectiva o emocional se está representando; en el protocolo de respuesta se señalan dos alternativas. Por ejemplo, un reactivo visual muestra el rostro de la mujer durante dos segundos mientras representa una escena emocional. Se pide al individuo examinado que elija entre dos descripciones de lo que está haciendo la persona: a) regañar a un niño o b) expresar enojo por celos. En este reactivo se "corta" el audio, de modo que la persona examinada tiene que decodificar la información a partir únicamente de la expresión facial. Dado que la mujer que hace la representación se encuentra a la vista durante sólo dos segundos, existe poca oportunidad de que el individuo utilice la lectura de labios para obtener información verbal útil.

El PONS evalúa la capacidad de la persona examinada para decodificar conducta no verbal a partir de 11 canales diferentes. Éstos incluyen tres que son canales visuales "puros", dos auditivos (de paralenguaje) diferentes y seis combinados (visuales y auditivos). Los visuales incluyen:

1. El rostro.
2. El cuerpo del cuello hasta las rodillas.
3. La figura completa.

Los canales auditivos incluyen los siguientes:

4. La voz empalmada al azar (EA) de la hablante.
5. La voz con filtración de contenido (FC) de la hablante.

Los canales combinados incluyen:

6. El rostro más la voz EA.
7. El cuerpo más la voz EA.
8. La figura más la voz EA.
9. El rostro más la voz FC.
10. El cuerpo más la voz FC.
11. La figura más la voz FC.

En los PONS, se muestra a la intérprete expresando 20 situaciones afectivas o emocionales diferentes que van desde emociones relativamente sutiles (p. ej., "expresar amor materno") hasta situaciones más dramáticas (p, ej., "amenazar a alguien"). Dado que cada una de las 20 escenas aparece en cada uno de los 11 canales, la prueba consiste de un total de 220 escenas. El PONS requiere cerca de 45 minutos para completarla.

La principal muestra de estandarización para el PONS consiste de 492 estudiantes de preparatoria pública de tres lugares (costa oeste, medio oeste y costa este) en EUA. Estos individuos tenían una inteligencia promedio y provenían principalmente de familias de clase media. Al utilizar la fórmula KR-20, los autores calcularon las correlaciones de consistencia interna de los resultados de prueba de estos estudiantes. Dichos resultados indican, en general, una homogeneidad adecuada para los canales individuales, con la evidente excepción del canal de voz empalmada al azar ($r = .06$). Por desgracia, se encontró que los coeficientes de estabilidad *test-retest* para las calificaciones de canal del PONS eran marginales cuando mucho (cuadro 14–12). Con base en 293 estudiantes a quienes se volvió a examinar después de 10 días a 10 semanas, la mayor parte de las correlaciones *test-retest* para las calificaciones de canales individuales se localizaron en los rangos de .20 y .30. El único coeficiente de estabilidad que resultó aceptable de manera marginal se basó en la escala completa ($r = .69$).

La validez del PONS se ha investigado en docenas de estudios que resumen los autores (Rosenthal y cols., 1979). En general, las correlaciones con otros índices de sensibilidad interpersonal son apropiadamente positivas (en apoyo de la validez de criterio), mientras que las correlaciones con variables cognitivas como el CI son adecuadamente bajas (en apoyo de la validez de discriminación). En un interesante estudio, se encontró que la eficiencia de los

* El PONS puede obtenerse en cinco versiones, incluyendo sólo audio, sólo vídeo, PONS breve y prueba de discrepancia no verbal. Aquí únicamente se analiza el PONS completo.

Cuadro 14–12. Coeficientes medianos de confiabilidad *test-retest* del PONS de seis muestras ($N = 293$)

Canal de audio	Canal de vídeo				
	Ninguno	Rostro	Cuerpo	Figura	Total
Ninguno		.24	.34	.24	.52
Empalmada al azar	.18	.43	.26	.20	.50
Con filtración de contenido	.27	.20	.24	.27	.50
Total	.32	.49	.54	.51	.69

Fuente: Reproducido con autorización de Rosenthal, R., Hall, J. A., DiMatteo, M. R., Rogers, P. L. y Archer, D. (1979). *Sensitivity to nonverbal communication: The PONS test.* Baltimore, MD: Johns Hopkins University Press.

funcionarios del Servicio Exterior se correlacionaba en .30 con las calificaciones de la versión de audio del PONS; es decir, los funcionarios más eficaces tuvieron mejores resultados que los menos eficientes en cuanto a decodificar la información contenida en las versiones de voz con filtración de contenido y de voz empalmada al azar. Por tanto, se puede concluir que la calificación total del PONS es un índice adecuado de la sensibilidad a la comunicación no verbal. Sin embargo, los canales individuales de la prueba poseen tal confiabilidad marginal que debería utilizárseles sólo con propósitos de investigación.

RESUMEN

1. La evaluación conductual se concentra en el comportamiento en sí, más que en los rasgos subyacentes, causas hipotéticas o supuestas dimensiones de la personalidad. En general, la evaluación conductual es una parte integral de la terapia conductual diseñada para cambiar la duración, frecuencia o intensidad de una conducta blanco bien definida.
2. Un enfoque útil de evaluación en los métodos de terapia conductual basados en la exposición es la *Behavioral avoidance Test* (BAT), en la que el terapeuta proyecta cuánto tiempo puede tolerar el cliente un estímulo inductor de ansiedad. También son útiles los *Fear Survey Schedules*, que se basan en las calificaciones de la persona para objetos y situaciones comúnmente temidos, pero existen razones para cuestionar su validez.
3. En el manejo de contingencias, el terapeuta intenta modificar la conducta desadaptada al identificar las consecuencias reforzantes y eliminarlas. A la inversa, las conductas adaptativas pueden fortalecerse al disponer que se presenten recompensas cuando éstas ocurren, como en la economía de fichas.
4. En la terapia cognitiva-conductual, el terapeuta intenta cambiar la estructura de creencias del cliente. Por ejemplo, Meichenbaum le enseña a los clientes a utilizar autoafirmaciones de enfrentamiento (p. ej., "Tienes una estrategia... puedes lograrlo") para combatir situaciones estresantes.
5. Un excelente índice de la depresión —incluyendo las distorsiones cognitivas— es

el *Beck Depression Inventory* (BDI), que consiste de 21 cuartetos de afirmaciones ordenadas de manera jerárquica, cada una de las cuales se califica de 0 a 3. Al BDI se le reconoce como un excelente índice de autoinforme de la depresión y se le ha validado ampliamente contra criterios externos.

6. Lewinsohn y sus colaboradores han publicado el *Pleasant Events Schedule* para el automonitoreo de la frecuencia y agradabilidad de hasta 320 conductas principalmente comunes y cotidianas. Los pacientes deprimidos que aumentan su frecuencia automonitoreada de acontecimientos agradables muestran, por lo general, una mejoría en su estado de ánimo.
7. El automonitoreo tiene muchas aplicaciones en psicología y medicina. Un ejemplo es el *Body Image Testing System* (BITS), que se utiliza para evaluar las distorsiones en imagen corporal que se encuentran en la anorexia nerviosa y otros trastornos de la alimentación. El BITS le proporciona al individuo un menú para cambiar una imagen corporal generada por computadora hasta que se parezca a su percepción de sí mismo.
8. Existen varias herramientas y métodos para la evaluación de conducta no verbal como la interacción visual. Por desgracia, la evaluación de ésta se halla llena de escollos prácticos y metodológicos; la aplicación de este enfoque se restringe a los estudios de investigación.
9. La paralingüística se refiere al tono de voz, tasa de verbalizaciones y otros aspectos no verbales del habla. Una importante área en la evaluación paralingüística ha sido la del juicio de emociones a partir del discurso con "filtración del contenido". Por ejemplo, el enojo paralingüístico percibido en la voz de un médico (juzgado en el discurso con filtración de contenido) predice una efectividad deficiente en la canalización de pacientes alcohólicos a tratamiento.
10. La expresión facial es otra conducta no verbal útil en la evaluación. El *Facial Action Coding System* (FACS) es un enfoque sistemático para la codificación de expresiones faciales a partir de videocintas en cámara lenta. El sistema consiste de 44 unidades de acción (UA) que abarcan todos los movimientos faciales visibles. Con el FACS, jueces entrenados pueden producir calificaciones sumamente confiables de las acciones faciales.
11. El *Profile of Nonverbal Sensitivity* (PONS) es una evaluación amplia de la capacidad para percibir señales no verbales en una serie de 220 segmentos de dos segundos (en película o videocinta) sobre la conducta no verbal de una mujer. Aunque la confiabilidad de los canales del PONS es baja, su calificación global es un índice adecuado de la sensibilidad a la conducta no verbal.

TÉRMINOS Y CONCEPTOS CLAVE

Evaluación conductual p. 645
Terapia conductual p. 647
Behavioral avoidance Test p. 648
Fear Survey Schedule p. 648
Manejo de contingencias p. 650
Economía de fichas p. 650
Terapia cognitiva-conductual p. 650
Automonitoreo p. 653
Conducta no verbal p. 655
Paralingüística p. 657

Capítulo 15

Temas y cuestiones especiales en las pruebas

Tema 15A
Evaluación por computadora y el futuro de las pruebas

Computadoras en el proceso de prueba: revisión y antecedentes
Interpretación de una prueba basada en sistemas: situación actual
Multimedia y realidad virtual: el nuevo horizonte de la CAPA
Valoración de la interpretación de pruebas basada en sistemas
Ejemplo de caso 15–1. Confianza irreflexiva en un informe del MMPI por computadora.
Pruebas adaptativas computarizadas
El futuro de las pruebas
Resumen
Términos y conceptos clave

En los capítulos anteriores se ha esbozado un sinnúmero de maneras en las que se utilizan las pruebas para la toma de decisiones. Además, se ha establecido que las pruebas psicológicas no sólo se han propagado sino que también tienen consecuencias. Los resultados de prueba son importantes. Las conclusiones de las pruebas pueden justificar el derecho a obtener algún privilegio. Al contrario, las conclusiones de una prueba pueden autorizar que se nieguen oportunidades. Entonces, por muchas razones, es adecuado concluir este libro con dos temas especiales relacionados con las repercusiones potenciales de las pruebas psicológicas. En el tema 15A, Evaluación por computadora y el futuro de las pruebas, se revisan las aplicaciones actuales de la computación en la evaluación psicológica y después se analizan las cuestiones profesionales y sociales que surgen de esta práctica. Este tema cierra con

las consideraciones acerca del futuro de las pruebas —que en gran medida se verán moldeadas por las aplicaciones cada vez más complejas de la tecnología informática. La materia de las cuestiones profesionales continúa en el tema 15B, Cuestiones éticas y sociales en las pruebas. Se comienza con una perspectiva general y antecedentes de la información en las pruebas.

COMPUTADORAS EN EL PROCESO DE PRUEBA: REVISIÓN Y ANTECEDENTES

INTRODUCCIÓN A LA EVALUACIÓN ASISTIDA POR COMPUTADORA

En muchos centros de orientación es posible que el cliente haga una cita con una microcomputadora para explorar sus opciones de carrera. Además de una breve interacción con la recepcionista para programar el horario de uso de la computadora, el cliente no necesita interactuar con ningún otro ser humano durante todo el proceso de evaluación. El escenario exacto diferirá de un ambiente a otro, pero podría parecerse al siguiente. Las instrucciones en la pantalla de la computadora alientan al usuario a oprimir cualquier tecla. Entonces, la computadora estimula al cliente a responder una serie de preguntas sobre actividades e intereses al oprimir teclas numéricas establecidas. Después de terminar con el inventario, la computadora calcula las puntuaciones naturales de una larga lista de escalas ocupacionales y realiza las transformaciones estadísticas apropiadas. A continuación aparece en pantalla un breve informe. Éste proporciona una lista de carreras que son las que más se adecuan a los intereses del cliente. También se imprime una copia para revisiones posteriores. Se puede suponer que el cliente tiene mejor información acerca de opciones compatibles de carrera y, por tanto, está en mayor probabilidad de elegir una profesión satisfactoria. Este escenario es un ejemplo simple de la evaluación psicológica asistida por computadora (CAPA; del inglés *computer-assisted psychological assesment*), un avance reciente bienvenido por muchos psicólogos pero también criticado por otros.

Es bien sabido que en la actualidad las computadoras tienen amplia utilización en las pruebas psicológicas. No obstante, la amplitud de estas aplicaciones podría sorprender al lector. Además de las aplicaciones directas como la presentación de preguntas de una prueba, calificación de los datos de la misma e impresión de resultados (como se dijo antes), las computadoras pueden utilizarse para: 1) diseñar pruebas individualizadas que se basen en realimentación en tiempo real durante la prueba; 2) interpretar los resultados de una prueba según reglas de decisión complejas; 3) redactar informes narrativos extensos y detallados y 4) presentar los estímulos de una prueba en un formato atractivo y realista de vídeo a través de multimedia. En ésta se tocaron todos estos temas. El término general **evaluación psicológica asistida por computadora** (CAPA) se refiere al rango completo de aplicaciones computacionales en la evaluación psicológica. La CAPA es muy prometedora para la práctica de la psicología, pero también presenta una variedad de problemas prácticos y éticos que demandan una consideración cuidadosa y atenta. Una historia breve sobre la CAPA sería un buen telón de fondo para el análisis de las preocupaciones prácticas y éticas.

ANTECEDENTES BREVES DE LA CAPA

La calificación manual de las pruebas psicológicas es tediosa, lenta y propensa al error. Por tanto, los psicólogos aceptaron fácilmente la tecnología en su búsqueda por mejorar la eficacia y precisión en las pruebas. El empleo de máquinas de calificación mecánica para las pruebas psicológicas, como los *Strong Vocational Interest Blanks* (SVIB), ocurrió por primera vez en el decenio de 1920. En 1946, Elmer Hankes construyó una computadora analógica para ob-

tener de manera automática las calificaciones y el perfil de la SVIB (Moreland, 1992). Para inicios del decenio de 1960, la combinación de lectores ópticos y computadoras centrales proporcionó una manera rápida y libre de errores de calificar e imprimir el perfil de pruebas como el SVIB y el MMPI.

La utilización de computadoras para obtener interpretaciones de una prueba —no sólo calificaciones y perfiles— puede rastrearse hasta la Clínica Mayo a principios del decenio de 1960 (Swenson, Rome, Pearson y Brannick, 1965). El grupo Mayo necesitaba un sistema rápido y eficaz para detectar los problemas psicológicos de miles de pacientes médicos a través del MMPI. Éstos respondían a los reactivos del MMPI en tarjetas especiales de la IBM que la computadora podía leer con un analizador. El primer sistema interpretativo era burdo si se le juzga por las normas actuales:

> Se escribió un programa que calificaba 14 escalas del MMPI, las convertía en puntuaciones estándar e imprimía una serie de afirmaciones descriptivas. Éstas se seleccionaban de un conjunto de 62 afirmaciones, la mayor parte de las cuales se asociaban con elevaciones en las escalas del MMPI. El programa tenía algunas afirmaciones sobre configuraciones, pero en su mayor parte se ignoraban las combinaciones de escalas, en las que se basa la mayor parte de la literatura (Fowler, 1985).

Las afirmaciones sobre configuraciones se refieren a interpretaciones basadas en patrones específicos de puntuaciones de escala; como las elevaciones más altas en dos escalas designadas. El éxito del sistema Mayo sirvió de ímpetu para la interpretación basada en computadora para muchas otras pruebas psicológicas. Por ejemplo, a principios del decenio de 1960, Piotrowski (1964) desarrolló un sistema de interpretación del Rorschach por computadora. El sistema para el Rorschach requería considerable "preprocesamiento" de parte de un técnico. Primero se codificaban las respuestas individuales según una lista de 320 parámetros. La interpretación se basaba en estas calificaciones de parámetro y no en las respuestas naturales.

Para el decenio de 1970, los psicólogos se dieron cuenta de que las computadoras podían integrarse en todo el proceso de evaluación psicológica. Johnson y Williams (1975) describieron el uso de una computadora central con varias terminales remotas para evaluar un promedio de 17 admisiones de internamiento psiquiátrico por día. Por lo común, los pacientes respondían a las siguientes pruebas aplicadas por computadora: MMPI, *Beck Depression Inventory*, prueba de inteligencia, prueba de memoria y antecedentes sociales. Un entrevistador llevaba a cabo un examen estructurado del estado mental e ingresaba los datos directamente a la computadora. Ésta calificaba las pruebas y generaba un informe narrativo completo. En una serie de estudios de investigación, el grupo Utah demostró que estos informes se generaban en la mitad del tiempo y a la mitad del costo de las evaluaciones tradicionales (Klingler, Miller, Johnson y Williams, 1977).

Para el decenio de 1980, la CAPA era tan frecuente que casi toda prueba psicológica existente podía interpretarse por computadora. Una cronología detallada de estos avances está más allá del espectro de este libro, pero en el cuadro 15–1 se han resumido los principales hitos históricos.

INTERPRETACIÓN DE UNA PRUEBA BASADA EN SISTEMAS: SITUACIÓN ACTUAL

La **interpretación de prueba basada en sistemas** (CBTI, del inglés *computer-based test interpretation*) se refiere a la interpretación de la prueba y redacción del informe por computadora. En la actualidad, las principales editoriales de pruebas ofrecen interpretaciones por computadora. Estos servicios están disponibles a través de correo, computadoras en línea con módem o paquetes de microcomputadora *on-site*. Lo que es más, el mercado de pruebas y de redacción de informes por computadora es tan lucrativo que se puede esperar un enorme crecimiento en este campo en los siguientes años. Butcher (1987, Apéndice A) lista 169 distribuidores comerciales para 1986. Conoley, Plake y

Cuadro 15–1. Hitos históricos de la CAPA

1946	Hankes desarrolla una computadora análoga para calificar el SVIB (Moreland, 1992).
1954	El libro *Clinical Versus Statistical Prediction*, de Meehl (1954), establece la pauta para la interpretación automatizada de las pruebas.
1962	El lector óptico y la computadora digital se utilizan para calificar el SVIB y el MMPI y también para imprimir perfiles. (Moreland, 1992)
1962	El primer sistema de interpretación de pruebas basado en sistemas para el MMPI se desarrolló en la Clínica Mayo (Swenson y cols., 1965).
1964	Piotrowski publica un sistema de interpretación por computadora para el Rorschach (Piotrowski, 1964).
Decenio de 1960	Proliferan los sistemas de interpretación por computadora para el MMPI; Fowler, Finney y Caldwell desarrollan sistemas populares (Fowler, 1985).
1971	Se utiliza una computadora central con terminales para automatizar todo el proceso de evaluación para pacientes internos en el área psiquiátrica del Hospital de Veteranos en Salt Lake City, Utah (Klingler, Miller, Johnson y Williams, 1977).
Decenio de 1970	Se introducen las pruebas adaptativas computarizadas (CAT); las CAT permiten baterías flexibles e individualizadas de prueba que producen un nivel dado de precisión de medida con el menor número posible de reactivos de una prueba. (Weiss, 1982)
1975	Primera interpretación automatizada de una batería de pruebas neuropsicológicas (Adams y Heaton, 1985).
1979	Lachar publica un sistema interpretativo con base actuarial para el *Personality Inventory for Children* (Lachar y Gdowski, 1979).
1985	Aparece una serie especial sobre evaluación psicológica por computadora en el *Journal of Consulting and Clinical Psychology* (Butcher, 1985).
1986	La *American Psychological Association* publica las *Guidelines for Computer-Based Tests and Interpretations.*
1987	Publicación del primer libro sobre recursos titulado: *Computarized Psychological Assessment: A Practitioner's Guide* (Butcher, 1987)
1994	Introducción a las baterías de prueba en multimedia; por ejemplo, en la IBM se utiliza una prueba en multimedia para evaluar las habilidades de solución de problemas reales de los posibles empleados (*Monitor* de la APA, junio de 1994).
1997	El *Educational Testing Service* y otros gigantes de las pruebas han cambiado a las pruebas por computadora para los principales exámenes de admisión, como la *Graduate Management Admission Test* (GMAT; Prueba de Administración de Admisiones para el Posgrado) y los *Graduate Record Examinations* (GRE; Exámenes del Registro de Posgrado).

Kemmerer (1991) señalan que el número de interpretaciones por computadora de pruebas psicológicas había aumentado a más de 400 para 1990. Casi cada mes se informa de nuevos sistemas de pruebas por computadora en las revistas y periódicos del ramo (p. ej., *Monitor* de la APA). Las interpretaciones de prueba basadas en sistemas llegaron para quedarse.

En esta sección se proporciona una revisión general de los tipos de interpretaciones de prueba basadas en sistemas que están disponibles en la actualidad. Una reseña completa de los productos podría abarcar con toda facilidad varios volúmenes, de modo que el lector tendrá que conformarse con un análisis de ejemplos diversos y representativos de la CBTI. Se revisarán cuatro aspectos de la CBTI: informes de calificación, informes descriptivos, informes actuariales e informes clínicos asistidos por computadora (Moreland, 1992).

INFORMES DE CALIFICACIÓN

Los informes de calificación expresan las puntuaciones, perfiles o ambos. Además, un informe de este tipo puede incluir la graficación de pruebas de significación estadística e intervalos de confianza para las calificaciones de una prueba. Por definición, los informes de calificación no incluyen texto o explicación narrativos acerca de las puntuaciones. Moreland (1992) analiza el atractivo de los informes de calificación:

> Estos tipos de datos hacen posible identificar a primera vista las calificaciones especialmente significativas y las diferencias importantes entre éstas. También deberían aumentar la confianza del usuario en que dichas calificaciones son de hecho importantes. Las pruebas de significación estadística son, sin duda, superiores a los "métodos clínicos prácticos", cuando se trata de la interpretación precisa de las calificaciones de una prueba. Y ¿quién tiene el tiempo para calcular los intervalos de confianza —en especial con pruebas que tienen docenas de escalas?

En la figura 15–1 se muesta como ejemplo un informe de calificación para el *Jackson Vocational Interest Survey* (Jackson, 1991). El lector observará que se presenta gran cantidad de información de una manera eficaz y condensada. Esto es típico de los informes de calificación. En una sola página, la persona hipotética que hubiera respondido a la prueba se enteraría de que sus intereses son sumamente similares a las licenciaturas en humanidades, educación y negocios. En términos de adaptación laboral, sabría también que es muy compatible con orientadores, maestros, abogados, administradores y otras profesiones que tienen énfasis en las relaciones humanas.

INFORMES DESCRIPTIVOS

Un informe descriptivo va un paso más allá del informe de calificación al proporcionar una interpretación breve, escala por escala, de los resultados de una prueba. Estos informes son especialmente útiles cuando las conclusiones de una prueba se comunican a profesionales de la salud mental que tienen poco conocimiento de la prueba en cuestión. Por ejemplo, la mayoría de los psicólogos clínicos saben que una alta calificación en la escala de Psicastenia del MMPI significa preocupación e insatisfacción con las relaciones sociales —pero otros profesionales de la salud mental pueden no tener idea del significado de una elevación en esta escala. Un informe descriptivo puede trasmitir información invaluable en media página o menos. En la figura 15–2 se muestra uno de los primeros informes descriptivos publicados. El lector verá que este paciente varón de 20 años de edad se describe como tímido, sensible, preocupado y gravemente deprimido. Es claro que está justificada la canalización de este paciente médico a un psicólogo o psiquiatra. Este informe es un modelo de sencillez y claridad. En comparación, la mayor parte de los informes descriptivos actuales que se basan en sistemas proporcionan detalles excesivos. Por lo común, el clínico debe batallar con la lectura de varias páginas de texto, a fin de extraer las características esenciales acerca del cliente.

INFORMES ACTUARIALES: PREDICCIÓN CLÍNICA *VERSUS* ACTUARIAL

El enfoque actuarial de la interpretación de una prueba basada en sistemas se basa en la determinación empírica de relaciones entre los resultados de una prueba y criterios de interés. La naturaleza de este enfoque puede comprenderse mejor en el contexto del antiguo debate entre la predicción clínica y la actuarial. Aquí es necesaria una breve desviación para introducir al lector a los conceptos y asuntos relacionados antes de analizar los informes actuariales.

Muchas interpretaciones de prueba basadas en sistemas hacen predicciones sobre la persona examinada. Con frecuencia, éstas se disfrazan en el lenguaje de la clasificación o el diagnóstico, pero siguen siendo predicciones. Por ejemplo, cuando un informe de prueba neuropsicológica por computadora expresa de manera tentativa un daño cerebral en un cliente, en realidad ésta es una predicción implícita

CASO EXAMINADO, M 123456789 VARÓN 27-MAR-84 PÁGINA 6

SEMEJANZA CON GRUPOS DE ESTUDIANTES UNIVERSITARIOS

	MUJERES	VARONES	
AGRICULTURA	-0.67	-0.57	(MUY BAJA)
ARTES Y ARQUITECTURA	-0.18	-0.08	(BAJA)
ADMINISTRACIÓN DE EMPRESAS	+0.53	+0.65	(MUY ALTA)
CIENCIAS DE LA TIERRA Y MINERALES		-0.72	(MUY BAJA)
EDUCACIÓN	+0.64	+0.65	(MUY ALTA)
INGENIERÍA	-0.53	-0.61	(MUY BAJA)
SALUD, EDUC. FÍSICA Y RECREACIÓN	-0.40		
DESARROLLO HUMANO	+0.39	+0.70	(MUY ALTA)
HUMANIDADES	+0.71	+0.77	(MUY ALTA)
CIENCIAS NATURALES	-0.66	-0.60	(MUY BAJA)
ENFERMERÍA	+0.17		
ESTUDIANTES DE MEDICINA	+0.08	-0.12	(MUY BAJA)
EDUCACIÓN TÉCNICA		-0.43	(MUY BAJA)

SEMEJANZA CON CLASIFICACIONES LABORALES

A CONTINUACIÓN SE CATALOGA LA CLASIFICACIÓN LABORAL QUE, SEGÚN SE ENCONTRÓ, ES SIMILAR A SU PERFIL DE INTERESES. UNA CALIFICACIÓN POSITIVA INDICA QUE SU PERFIL MUESTRA CIERTO GRADO DE SEMEJANZA CON AQUELLOS QUE YA TRABAJAN EN EL GRUPO LABORAL, MIENTRAS QUE UNA NEGATIVA INDICA FALTA DE SEMEJANZA.

CALIFICACIÓN	SEMEJANZA	CLASIFICACIÓN LABORAL
+0.78	MUY SEMEJANTE	ORIENTACIÓN/ESTUDIANTES PERSONAL TRABAJADORES
+0.74	MUY SEMEJANTE	ENSEÑANZA Y OCUPACIONES RELACIONADAS
+0.74	MUY SEMEJANTE	OCUPACIONES EN RELIGIÓN
+0.71	MUY SEMEJANTE	ADMINISTRACIÓN Y OCUPACIONES RELACIONADAS
+0.68	MUY SEMEJANTE	OCUPACIONES EN LEYES Y POLÍTICA
+0.60	MUY SEMEJANTE	ADMINISTRACIÓN DE PERSONAL/RECURSOS HUMANOS
+0.57	SIMILAR	OCUPACIONES EN BIENESTAR SOCIAL
+0.55	SIMILAR	OCUPACIONES EN CIENCIAS SOCIALES
+0.55	SIMILAR	OCUPACIONES EN ENSEÑANZA PREESCOLAR Y PRIMARIA
+0.50	SIMILAR	OCUPACIONES EN VENTAS
+0.50	SIMILAR	OCUPACIONES EN MERCADEO
+0.49	SIMILAR	SERVICIOS SECRETARIALES
+0.48	SIMILAR	LABORES DE REDACCIÓN
+0.44	SIMILAR	OCUPACIONES EN CONTADURÍA, BANCA Y FINANZAS
+0.05	NEUTRO	OCUPACIONES DE SERVICIO
+0.03	NEUTRO	OCUPACIONES EN LA MÚSICA
+0.02	NEUTRO	OCUPACIONES DE ENSAMBLADO-INSTRUMENTOS Y PEQUEÑOS PRODUCTOS

Figura 15–1. Un informe de calificación para la Encuesta Jackson de interés vocacional.
Fuente: Reproducido con autorización de Jackson, D. N. (1991). *Jackson Vocational Interest Survey Manual* (3a. ed.). Port Huron, MI: Sigma Assessment Systems, Inc. (800) 265-1285.

CALIFICACIÓN	SEMEJANZA	CLASIFICACIÓN LABORAL
-0.30	NEUTRO	OCUPACIONES EN EL ESPECTÁCULO
-0.24	NEUTRO	OCUPACIONES EN ARTES PUBLICITARIAS
-0.35	DIFERENTE	OCUPACIONES EN SERVICIOS DE PROTECCIÓN
-0.38	DIFERENTE	AGRICULTURA
-0.39	DIFERENTE	OFICIALES MILITARES
-0.41	DIFERENTE	OCUPACIONES EN BELLAS ARTES
-0.58	DIFERENTE	OCUPACIONES EN DEPORTES Y RECREACIÓN
-0.59	DIFERENTE	MATEMÁTICAS Y LABORES RELACIONADAS
-0.62	MUY DIFERENTE	MAQUINARIA/MECÁNICA Y TRABAJOS RELACIONADOS
-0.63	MUY DIFERENTE	TRABAJOS EN SERVICIOS DE SALUD
-0.65	MUY DIFERENTE	OCUPACIONES EN CIENCIAS FÍSICAS
-0.65	MUY DIFERENTE	CONSTRUCCIÓN/OBREROS CALIFICADOS
-0.68	MUY DIFERENTE	OCUPACIONES EN DIAGNÓSTICO Y TRATAMIENTO MÉDICO
-0.71	MUY DIFERENTE	TRABAJADORES EN INGENIERÍA Y APOYO TÉCNICO
-0.76	MUY DIFERENTE	CIENCIAS BIOLÓGICAS

Figura 15–1. (continuación).

que puede confirmarse o descartarse mediante criterios externos como una exploración cerebral o una consulta neurológica. De la misma manera, cuando un informe del MMPI-2 basado en sistemas proporciona un diagnóstico tentativo de canalización clínica, según criterios del DSM-IV, ésta también es una predicción que puede validarse o invalidarse a través de criterios externos como una entrevista clínica intensiva. Un ejemplo final: cuando un informe de selección CPI por sistemas para candidatos a ingresar a la policía advierte que un solicitante tendrá una adaptación deficiente dentro del sistema judicial, esto también constituye una predicción que puede resultar correcta o incorrecta a través de una inspección de los registros de personal en una fecha posterior.

El empleo de computadoras para la predicción basada en pruebas destaca una distinción esencial que se conoce como juicio clínico *versus* juicio actuarial (Dawes, Faust y Meehl, 1989; Garb, 1994; Meehl, 1954; 1965, 1986). En el **juicio clínico**, la persona que toma la decisión procesa la información dentro de su cabeza para diagnosticar, clasificar o predecir la conducta. Un ejemplo es el psicólogo clínico que utiliza su experiencia, intuición y conocimientos académicos para determinar si un perfil del MMPI indica o no la presencia de psicosis. Ésta constituye una categoría amplia que incluye graves trastornos mentales que con frecuencia se caracterizan por alucinaciones, delirios y trastornos del pensamiento. Así, la predicción (o no) de psicosis de parte de un clínico puede validarse contra criterios externos como entrevistas detalladas.

En el **juicio actuarial**, se utiliza una fórmula derivada empíricamente para diagnosticar, clasificar o predecir la conducta. Un ejemplo es el psicólogo clínico que sólo ingresa las calificaciones de una prueba en una fórmula basada en la investigación para determinar si un perfil del MMPI indica o no la presencia de psicosis. La predicción actuarial también puede validarse contra criterios externos adecuados.

La esencia del juicio actuarial es el desarrollo cuidadoso y posterior utilización de una fórmula con base empírica para el diagnóstico, clasificación o predicción de la conducta. Un tipo común de fórmula actuarial es la ecuación de regresión, en la que las calificaciones de una subprueba se combinan en una suma lineal ponderada para predecir un criterio relacionado. Pero otros enfoques estadísticos también pue-

Sexo: Masculino Educación: 20 Edad: 34 años Estado Civil: Casado Paciente externo

Clave del MMPI: 27'5'8064-391/-KLF/

D	2	Gravemente deprimido, preocupado, indeciso, pesimista
Pt	7	Rígido y meticuloso. Inquieto y aprensivo. Insatisfecho con relaciones sociales. Probablemente muy religioso y moralista
Mf	5	Probablemente sensible e idealista con elevados intereses estéticos, culturales y artísticos
Es	8	Tiende hacia intereses abstractos como la ciencia, filosofía y religión
Is	0	Probablemente retraído y tímido en situaciones sociales
Pa	6	Sensible. Atento a las opiniones de los demás
Dp	4	Independiente y un tanto inconformista
Hi	3	
Ma	9	Nivel normal de energía y actividad
Hs	1	Número de síntomas físicos y preocupación con funciones corporales bastante típico de pacientes clínicos

Considérese una valoración psiquiátrica

Figura 15–2. Informe descriptivo para el MMPI de la Clínica Mayo.
Fuente: Reproducido con autorización de Dahlstrom, W. G., Welsh, G. S. y Dahlstrom, L. E. (1972). *An MMPI handbook. Volume 1: Clinical interpretation* (ed. rev.). Minneapolis, MN: University of Minnesota Press. Copyright © 1960, 1972 de la Universidad de Minnesota.

den funcionar de manera apropiada para la toma de decisiones, incluyendo simples puntuaciones de corte y diagramas de flujo basados en reglas. Por supuesto, las reglas estadísticas se prestan para ponerlas en práctica en una computadora, de modo que es adecuado analizar el juicio clínico *versus* el actuarial en esta sección sobre interpretación por computadora.

Aunque las computadoras facilitan el uso del método actuarial, es necesario enfatizar que "actuarial" y "computarizado" no son sinónimos. Para ser realmente actuarial, la interpretación de una prueba debe ser automática (especificada de antemano o con base en una rutina) y debe fundarse en relaciones establecidas a nivel empírico (Dawes, Faust y Meehl, 1989). Si un programa de cómputo incorpora dichas reglas automáticas, con base empírica, para la toma de decisiones, entonces se está haciendo una predicción actuarial. Por el contrario, si un programa de cómputo incorpora el pensamiento y juicio de un clínico —no importa qué tan sabia pueda ser esa persona— entonces se está realizando una predicción clínica.

Meehl (1954) fue el primero en presentar la cuestión del juicio clínico *versus* el actuarial ante un amplio grupo de científicos sociales. Este autor expresó la cuestión con una sencillez pura: "¿Cuándo debemos utilizar nuestra cabeza en vez de la fórmula?". Considérese este problema práctico con base en los resultados del MMPI. La **Neurosis** es un término diagnóstico anticuado (aunque todavía se utiliza) que se refiere a una forma más leve de trastorno mental en el que predominan síntomas de ansiedad y disforia. Como se señaló antes, la psicosis es una forma más grave de trastorno mental que puede incluir alucinaciones, delirios y trastornos del pensamiento. El diagnóstico diferencial entre estas dos clases amplias de trastorno mental es importante. Con frecuencia, las personas con neurosis responden bien a la psicoterapia individual, mientras que un paciente con psicosis puede requerir de poderosos fármacos antipsicóticos que producen efectos secundarios adversos. ¿Cuál es superior para la toma de decisiones diagnósticas con base en el MMPI, la mente de un psicólogo bien entrenado o una fórmula apropiada basada en investigación previa? Más adelante se regresará a este asunto.

Meehl (1954) especificó dos condiciones para una comparación justa entre estos dos enfoques contrastantes para la toma de decisiones. Primero, ambos métodos deben basar los juicios en los mismos datos. Por ejemplo, cuan-

do se compara al clínico experimentado contra una ecuación actuarial, ambos enfoques deberían hacer pronósticos para el mismo conjunto de perfiles del MMPI y para ésos. En segundo lugar, se deben evitar las condiciones que puedan inflar de manera artificial la precisión del enfoque actuarial. Por ejemplo, la ecuación actuarial debe obtenerse de una muestra inicial, antes de la comparación con la toma de decisiones por métodos clínicos en una nueva muestra de perfiles del MMPI. De otro modo, las reglas de decisión actuarial se beneficiarán de las relaciones aleatorias entre variables y producirán una tasa falsamente alta de decisiones correctas.

Cuando se satisfacen las condiciones para una comparación justa entre los enfoques clínico y actuarial en la toma de decisiones, el último método resulta superior en casi todos los casos. El enfoque actuarial es claramente mejor para la tarea citada antes —el diagnóstico diferencial entre neurosis y psicosis a partir del MMPI. L. R. Goldberg (1965) determinó que una simple suma lineal de las calificaciones de escala del MMPI resultaba en 70% de clasificaciones correctas, mientras que los psicólogos con doctorado promediaron sólo 62%, donde el mejor psicólogo entre todos ellos logró 67% de decisiones correctas. La regla de decisión que derrotó a todos los contendientes humanos fue: si la suma de puntuaciones *T* en L + Pa + Es – Hi – Pt supera a 44, entonces diagnosticar psicosis; de otra manera diagnosticar neurosis.*

Dawes, Faust y Meehl (1989) citaron cerca de 100 estudios comparativos en las ciencias sociales. En casi todos los casos, el método actuarial igualó o superó al método clínico, en ocasiones de manera sustancial. La investigación de Leli y Filskov (1984) es típica de este tipo de resultados. Estos autores estudiaron el diagnóstico de disfunción cerebral progresiva con base en pruebas neuropsicológicas. Se aplicó una regla de decisión actuarial obtenida de un conjunto de casos a una nueva muestra con 83% de identificación correcta. Al trabajar precisamente con los mismos datos de prueba, grupos de clínicos inexpertos y expertos identificaron de manera correcta sólo a 63 y a 58%, respectivamente, de los nuevos casos. El lector verá el hecho perturbador y vergonzoso de que la experiencia no mejoró las tasas de éxito para esta tarea de toma de decisiones por métodos clínicos.

La lección a aprender de estos trabajos es que los informes narrativos computarizados deberían incorporar métodos actuariales, siempre que sea posible. Por ejemplo, los informes generados por computadora deberían emplear las fórmulas actuariales existentes para determinar la probabilidad de diversos diagnósticos psiquiátricos, en vez de depender de la lógica programada de un clínico experto. Por desgracia, como el lector descubrirá a continuación, la mayor parte de los informes narrativos por computadora tienen una base clínica —lo cual hace surgir dudas sobre su validez.

INTERPRETACIÓN ACTUARIAL: ENFOQUE DE MUESTRA

Sines (1966) definió la interpretación actuarial como aquella que se basa en "la determinación empírica de las regularidades que pueden existir entre los datos específicos de una prueba psicológica y características importantes, externas a la prueba, detalladas de manera clara en términos sociales, clínicos o teóricos acerca de la persona sometida a prueba". En otras palabras, las afirmaciones en una interpretación actuarial no se obtienen de las conjeturas o saber clínico; se basan en las conclusiones específicas y cuantificadas de la investigación.

Debido al esfuerzo de investigación que se requiere, los enfoques actuariales en la interpretación de pruebas basadas en sistemas son poco comunes. Los primeros sistemas actuariales se basaron en el MMPI (p. ej., Gilberstadt y Duker, 1965; Marks y Seeman, 1963). Recientemente, se han aplicado sistemas de interpretación actuarial al *Personality Inventory for Children*; PIC (Inventario de Personalidad para Niños;

* En forma respectiva, los nombres completos de estas escalas son L (escala de validez), Paranoia, Esquizofrenia, Histeria y Psicastenia.

Lachar y Gdowski, 1979; Lachar, 1987), al *Marital Satisfaction Inventory* (Snyder, Lachar y Wills, 1988) y al Inventario psicológico de California (Gough, 1987). Un ejemplo específico aclarará este método.

Los autores del *Personality Inventory for Children* (PIC) han producido un sistema ejemplar para la interpretación actuarial por computadora que se describirá con propósitos ilustrativos. El lector recordará del capítulo anterior que el PIC es un inventario de cierto-falso que responde uno de los padres, o la persona encargada del menor, acerca de la conducta del niño. Con base en estas respuestas, se produce un perfil de puntuaciones *T* (media de 50, DE de 10) para cuatro escalas de validez (p. ej., Defensividad), 12 escalas clínicas (p. ej., Delincuencia) y cuatro escalas factoriales (p. ej., Incompetencia social). En total, se hace un informe de puntuaciones *T* para 20 escalas del PIC. Por supuesto, aquéllas más altas indican mayor probabilidad de psicopatología.

La interpretación actuarial del PIC depende de correlaciones empíricamente derivadas entre las escalas individuales y criterios importantes ajenos a la prueba. Los individuos de investigación del estudio de Lachar y Gdowski (1979) eran 431 niños canalizados a una atareada clínica educativa. Como parte del proceso de evaluación de cada niño, los miembros del personal, padres y maestros respondieron a un amplio cuestionario que listaba 322 afirmaciones descriptivas relacionadas con la conducta y otras variables. Además, los padres o encargados respondieron al PIC.

En la primera fase del estudio actuarial, se correlacionaron las 322 afirmaciones descriptivas con las 20 escalas del PIC para identificar correlatos significativos por escala. En una segunda fase, se analizaron a mayor profundidad los correlatos significativos para determinar la relación entre las afirmaciones descriptivas y los rangos de puntuación *T* en las escalas del PIC. El resultado de este esfuerzo prodigioso fue una serie de tablas actuariales no muy diferentes de aquellas que utilizan las compañías de seguros para predecir la probabilidad de enfermedad, muerte, accidentes y similares con base en datos demográficos de población, como edad, sexo y lugar de residencia. En el cuadro 15–2 se presentan algunos ejemplos de los correlatos actuariales de la Escala de Delincuencia o DLQ.

Las tablas actuariales captan un cúmulo de información útil para la práctica clínica. Considérense a dos niños hipotéticos de 12 años de edad,

Cuadro 15–2. Tasas de ocurrencia de descriptores actuariales de la Escala de delincuencia del PIC

Descriptor	Tasa base*	Rangos de puntuación *T*							
		30-59	60-69	70-79	80-89	90-99	100-109	110-119	>120
Se niega a ir a la cama	30	18	26	23	33	36	33	42	38
Miente	62	44	36	48	73	71	79	90	91
Utiliza sustancias prohibidas	12	0	2	6	7	11	18	32	53
Rechaza la escuela	40	16	26	40	42	50	47	56	67
Involucrado con la policía	17	0	4	6	10	21	19	58	63

* Porcentaje de todos los niños en los que se calificó la presencia de la característica.
Nota: Estos cinco descriptores sólo son una muestra representativa de los 51 correlatos actuariales de la Escala de delincuencia.
Fuente: Material del *Actuarial Assessment of Child and Adolescent Personality: An Interpretive Guide for the Personality Inventory for Children Profile* copyright © 1979 de Western psychological Services. Reproducido con autorización del editor, Western Psychological Services, 12031 Wilshire Boulevard, Los Angeles, CA 90025, EUA.

Jimmy y Johny, cada uno de ellos canalizado a un clínico por el mismo problema: bajo rendimiento escolar. Como parte del procedimiento inicial, el clínico le pide a cada una de las madres que responda al PIC. Supóngase que la calificación de Jimmy en la Escala de Delincuencia o DLQ está sumamente elevada, con una puntuación *T* de 114, mientras que John obtiene una puntuación *T* de 54, dentro del rango promedio. Con base en estas calificaciones, el clínico sabría la probabilidad —listada a continuación como porcentajes— de que ciertas descripciones conductuales se apliquen a cada niño:

	Jimmy (DLQ = 114)	Johny (DLQ = 54)
Se niega a ir a la cama	42%	18%
Miente	90%	44%
Utiliza drogas	32%	0%
Rechaza la escuela	56%	16%
Problemas con la policía	58%	0%

El lector reconocerá de inmediato que Jimmy encaja en un patrón de trastorno conductual generalizado, mientras que Johny parece tener pocos problemas de conducta. En el caso del primero, es más probable que el bajo rendimiento sea secundario a un patrón de conducta antisocial, mientras que en el caso de Johny, el clínico debe buscar en otra parte para comprender el fracaso en la escuela. Por supuesto, ésta es sólo una pequeña fracción de la información que se tendría disponible de una interpretación actuarial por computadora para el PIC. En un informe completo, el clínico recibiría los datos estadísticos y las afirmaciones narrativas pertinentes para las 20 escalas completas del PIC.

INFORMES CLÍNICOS ASISTIDOS POR COMPUTADORA

En un informe clínico asistido por computadora, las afirmaciones interpretativas asignadas a los resultados de una prueba se basan en el juicio de uno o más clínicos expertos. Éstos formalizan sus procesos de pensamiento y desarrollan reglas automatizadas de decisión que después se traducen a un código de computadora. Este método difiere de manera crucial del enfoque actuarial asistido por computadora, en el que las afirmaciones interpretativas se basan estrictamente en las conclusiones de investigación formal. A nivel superficial, los dos enfoques pueden parecer idénticos, en el sentido de que cada uno se basa en reglas y es automático. La diferencia tiene que ver con el origen de las reglas: investigación empírica (enfoque actuarial) *versus* juicio clínico (enfoque clínico).

Aunque en general los clínicos reconocen la superioridad del método actuarial, existe una ventaja significativa del enfoque clínico asistido por computadora. Ésta es que el enfoque clínico se puede diseñar para interpretar todos los perfiles de una prueba, mientras que algunos serán imposibles de interpretar por medio del enfoque actuarial. La desalentadora verdad acerca de los sistemas actuariales, tipo "libro de cocina", dirigidos a la interpretación de pruebas es que, en general, la tasa de clasificación desciende de manera vertiginosa cuando se utiliza un sistema en un nuevo entorno. La tasa de clasificación se refiere al porcentaje de resultados de una prueba que encaja en las complejas reglas de clasificación del perfil necesarias para la interpretación actuarial. Por ejemplo, en el sistema actuarial para el MMPI de Gilberstadt y Duker (1965), la clave de perfil 1-2-3 se define por las siguientes reglas para las escalas Hs (Hipocondriasis), D (Depresión), Histeria (Hi) y L, F, K (validez):

1. Hs, D y Hi por encima de $T = 70$
2. Hs > D > Hi
3. Ninguna otra escala por encima de $T = 70$
4. L < *T* de 66, F < *T* de 86 y K < *T* de 71.

Es frecuente que las personas que producen este tipo de perfil en el MMPI sufran de sobrerreac-tividad psicofisiológica, por no mencionar una multitud de otras características confirmadas a nivel empírico. Por supuesto, existen

varias claves de perfil adicionales, cada una de ellas defi-nida por un conjunto de reglas complejas de decisión y acompañadas por una descripción elaborada, con base actuarial, de la personalidad y psicopatología. Una conclusión típica es que un sistema actuarial asistido por compu-tadora que se desarrolla dentro de una población de clientes será capaz de interpretar hasta 85% de los perfiles de una prueba que se encuentren en ese entorno. No obstante, cuando el sistema actuarial se aplica a un nueva población de clientes, quizá 50% de los perfiles de una prueba encajarán en las reglas de decisión. Esto significa que cerca de la mitad de dichos perfiles no encajan en las reglas. Cuando mucho, estos clientes recibirán una interpretación superficial, escala por escala, en vez de una interpretación actuarial más compleja basada en las claves de perfil. En casi todos los estudios sobre la interpretación actuarial se observa el problema de la reducción en la tasa de clasificación (Moreland, 1992).

Los informes clínicos asistidos por computadora tienden a ser extensos y detallados, llenos de calificaciones de escalas, índices de reactivos y gráficas. Por supuesto, estos informes incluyen varias páginas de informe narrativo, que en general se frasea en términos de hipótesis en vez de conclusiones confirmadas. El informe más corto de este tipo es de cerca de seis páginas (p. ej., el *Karson Clinical Report* [Informe Clínico Karson] para el 16FP), mientras que los más grandes pueden constar de 10 o 20 páginas (p. ej., las interpretaciones para el MMPI-2).

MULTIMEDIA Y REALIDAD VIRTUAL: EL NUEVO HORIZONTE DE LA CAPA

Con los recientes avances en tecnología, la microcomputadora moderna ha abierto un mundo nuevo para la evaluación psicológica. La computadora común de escritorio es capaz ahora de presentar segmentos de vídeo que poseen la claridad visual de la televisión. Los sistemas estereo interconstruidos producen una salida exquisita de audio, incluyendo el habla humana sintetizada que parece verdadera. Con los accesorios de CD-ROM, es posible el acceso instantáneo a enormes almacenes de información —incluyendo imágenes fijas, segmentos de vídeo en vivo, música, tablas, gráficas, animación. En forma colectiva, estas capacidades se conocen como **multimedia**— en especial cuando se les emplea en aplicaciones interactivas y educativas. El multimedia es un nuevo horizonte de la evaluación psicológica asistida por computadora. El otro es la realidad virtual, que se analizará más adelante.

En la IBM, los investigadores han desarrollado una prueba computarizada en multimedia que puede emplearse para evaluar a los solicitantes de trabajo para puestos de manufactura (*Monitor* de la APA, junio de 1994). El aspecto único de esta prueba es la naturaleza de los estímulos. En vez de sólo describir una situación de trabajo, la computadora con multimedia representa la escena en sí. A medida que el solicitante observa las diferentes viñetas, la computadora congela la imagen en momentos cruciales y le pregunta al candidato qué debería hacer en esa situación. Las escenas proporcionan una sensación sumamente realista, lo cual aumenta la validez de facie de la prueba. Los sistemas multimedia ofrecen proporcionar una evaluación más precisa del desempeño real de las personas en el puesto que las pruebas de lápiz y papel. Las pruebas en multimedia son especialmente buenas para detectar la capacidad de la persona examinada para enfrentarse con problemas complejos y realistas, como la toma de decisiones bajo presión de tiempo o la solución de conflictos en el trabajo.

Otra aplicación potencial de los sistemas multimedia es la selección de personal para ingreso como agentes de la policía. El personal que se encarga del cumplimiento de la ley debe tener capacidad de observación y valoración, que pueden evaluarse de manera realista con estímulos en vídeo. Por ejemplo, una evaluación podría consistir, en parte, de cintas de vídeo acerca de testigos en una escena delictiva. Se podría pedir a los candidatos a ingresar a la

policía que determinaran la veracidad de los testigos y que extrajeran conclusiones acerca del delito con base en sus poderes de observación (*Monitor* de la APA, junio de 1994). Este ejemplo —hipotético en la actualidad— ilustra el potencial de multimedia para revolucionar la evaluación psicológica.

Por último, debería mencionarse que las pruebas en multimedia son casi independientes del lenguaje. Los candidatos talentosos a un empleo, que no poseen buenas capacidades de lectura o escritura, pero que tienen habilidades prácticas para el puesto, pueden identificarse por pruebas en multimedia. Para algunos empleos, los sistemas multimedia pueden ser más justos que el enfoque de lápiz y papel.

Por último, un enfoque muy reciente de tecnología avanzada para la evaluación con base en computadoras merece un comentario breve. En la realidad virtual, el participante utiliza un par de gafas que transmiten imágenes realistas, tridimensionales, de un ambiente simulado. Mediante la manipulación de mecanismos sencillos de control, el participante puede navegar a través del ambiente aunque permanezca en un solo sitio. Por supuesto, el ambiente visual, conocido como **realidad virtual**, se basa en complejos estímulos por computadora. Las nuevas herramientas de evaluación que utilizan la realidad virtual se encuentran en sus primeros años, pero parecen muy prometedoras. Por ejemplo, es posible valorar la imagen corporal de un cliente por medio de la realidad virtual, como se describe en un reciente artículo, en una nueva revista intitulada *Cyber Psychology and Behavior*.

VALORACIÓN DE LA INTERPRETACIÓN DE PRUEBAS BASADA EN SISTEMAS

Las pruebas por computadora tienen claras ventajas, pero también algunas desventajas potencialmente graves en comparación con el enfoque clínico tradicional de pruebas psicológicas. Aquí se ofrece una breve revisión que destaca tanto las ventajas como las desventajas de las pruebas, diagnóstico y redacción del informe por computadora. En Butcher (1987), Kramer (1988), Moreland (1992), Roid y Johnson (1998), Tallent (1987) y el número especial del *Journal of Consulting and Clinical Psychology*, diciembre (1985), pueden encontrarse mayores detalles acerca de este tema.

VENTAJAS DE LAS PRUEBAS Y REDACCIÓN DEL INFORME POR COMPUTADORA

Las principales ventajas de las pruebas por computadora son su rápido procesamiento, bajo costo, casi perfecta confiabilidad y completa objetividad. Además, algunas aplicaciones de medición, como las pruebas flexibles ajustables, requieren del uso de las computadoras para ponerlas en práctica. Después se exploran estos puntos con mayor detalle.

En una práctica clínica muy solicitada, las demoras entre la aplicación de la prueba y la presentación del informe de consultoría son comunes y casi inevitables. Éstas no sólo deterioran la reputación del psicólogo, sino que también pueden afectar de manera adversa el resultado del tratamiento para el cliente. Por ejemplo, un estudiante universitario con una discapacidad para el aprendizaje puede requerir una intervención inmediata, con la finalidad de evitar un desastre académico. Una demora de dos o tres semanas para presentar un informe de consultoría podría afectar, de manera indirecta, la diferencia entre el fracaso y el éxito en el desempeño académico. Los informes por computadora pueden acelerar todo el proceso de consultoría. Muchos sistemas de *software* producen informes que pueden transferirse a un programa de procesamiento de textos para la edición inmediata, personalizada, con lo cual se acelera el tiempo de procesamiento (p. ej., *Psychological Corporation*, 1994; Tanner, 1992).

El costo es otra consideración en las pruebas por computadora. Aunque no existen estudios definitivos sobre este tema, la mayoría de las autoridades afirman que las pruebas psicológicas calificadas e interpretadas por compu-

tadora cuestan considerablemente menos que aquéllas producidas completamente a través del esfuerzo de un clínico (Butcher, 1987). En sus estudios sobre pruebas automatizadas en el Hospital de Veteranos en Salt Lake City, Klinger, Miller, Johnson y Williams (1977) concluyeron que la computadora redujo el costo de las pruebas a la mitad. Con toda seguridad, a medida que los programas computarizados de pruebas se vuelvan más elaborados y los utilice un mayor número de clínicos, el costo por consulta disminuirá de manera notable.

La confiabilidad y objetividad constituyen el sello distintivo de las computadoras. Si se supone que el *software* es preciso y está libre de errores, las computadoras simplemente no pueden cometer errores de calificación, como tampoco pueden variar sus métodos de presentación del estímulo de un día para otro, ni producir informes narrativos diferentes con base en la misma información ingresada. El producto es el mismo sin importar las veces que se emplee el programa de cómputo. Lo que es más, dado que los informes por computadora se basan en reglas objetivas, no los distorsionan los efectos de halo u otros elementos de sesgo subjetivo que podrían afectar un informe obtenido por métodos clínicos. Butcher (1987) destaca que los informes por computadora pueden tener una importancia especial en casos judiciales, dado que se considera que "no los ha tocado la mano del hombre". Ésta es una posibilidad fascinante, pero quizá demasiado optimista en cierto sentido. Los abogados y jueces seguirán teniendo el deseo de saber quién programó el *software*, cómo se desarrollaron las afirmaciones narrativas y demás.

DESVENTAJAS DE LAS PRUEBAS Y REDACCIÓN DEL INFORME POR COMPUTADORA

Considérese el siguiente ejemplo, hipotético aunque realista y que probablemente ocurre en muchas ocasiones. El médico de un hospital canaliza a un paciente médico difícil al servicio psicológico para que se haga una evaluación de personalidad. Se lleva al paciente al centro de pruebas, donde una recepcionista lo sienta en una mesa frente a una microcomputadora. Las instrucciones que aparecen en el monitor de ésta piden que se responda a una serie de autoafirmaciones ciertas o falsas al oprimir la tecla C o F. El paciente responde al inventario objetivo de personalidad y se le regresa al servicio médico. Segundos después, de la impresora surge un informe narrativo basado en las respuestas del paciente. El psicólogo consultor revisa de manera breve el informe y después lo envía (sin firmar) al médico a través de correo interno. El informe es atractivo, con una apariencia precisa por su impresión en láser y con un resumen gráfico de las escalas en la primera página. Lo que es más, la narrativa suena válida y se ve como si la hubiera corregido un escritor profesional (de hecho, así fue). El médico se siente impresionado y adopta el informe como la verdad absoluta, tomando decisiones de tratamiento que se basan en la evaluación de personalidad.

Este escenario ilustra un dilema esencial de las pruebas y redacción de informes por computadora: éstas pueden dominar de tal manera el proceso de prueba que se relega al psicólogo clínico a un papel de mero recepcionista —o se le elimina por completo del proceso de evaluación. Aunque la mayoría de los psicólogos reconocen que las computadoras son una adición bienvenida a la práctica de las pruebas psicológicas, los críticos han presentado una serie de dudas inquietantes acerca de las prácticas recientes de evaluación como la presentada antes. La mecanización del proceso de prueba debido al uso de computadoras hace surgir cuestiones prácticas, legales, éticas e importantes que merecen una revisión cuidadosa.

En general, los escépticos no atacan la práctica de procesar por computadora la aplicación y calificación mecánicas de las pruebas; estas aplicaciones se consideran como usos eficientes y apropiados de la tecnología moderna. No obstante, incluso los defensores más fervientes reconocen la necesidad de investigar la equivalencia prueba-forma cuando se adapta una prueba existente a su aplicación por compu-

tadora. En particular, los profesionales no deberían suponer que la adaptación en computadora y la versión original de una prueba producen resultados idénticos. La equivalencia es una cuestión empírica que debe demostrarse con la investigación apropiada. En el caso de la mayor parte de las pruebas, puede demostrarse la equivalencia, pero éste no es un asunto que deba darse por sentado (Lukin, Dowd, Plake y Kraft, 1985; Schuldberg, 1988).

Algunas pruebas no conservan equivalencia de calificaciones cuando se les traduce a un sistema de cómputo. La Prueba de Categorías (*Category Test*; CT), de la *Halstead-Reitan Neuropsychological Battery*, es un caso a propósito. En una comparación de las versiones por computadora y estándar de la Prueba de categorías con pacientes en rehabilitación, Berger, Chibnall y Gfeller (1994) encontraron una enorme diferencia en la tasa de errores en dos grupos de personas que tenían antecedentes equivalentes: un promedio de 84 errores para la CT por computadora contra un promedio de 66 errores en la CT estándar. En apariencia la CT por computadora es mucho más difícil que la versión estándar, lo cual significa que deben desarrollarse normas independientes para su interpretación. También se han informado diferencias mucho más pequeñas entre las versiones por computadora y estándar del MMPI, donde las calificaciones de la versión por computadora tienden a subestimar (ligeramente) aquéllas basadas en el protocolo normal (Watson, Thomas y Anderson, 1992).

El punto de atención principal de la controversia sobre interpretación de pruebas por computadora es la redacción computarizada del informe. Varios expertos destacados en pruebas psicológicas han expresado graves dudas acerca de la práctica rutinaria de automatizar los informes narrativos que deben acompañar a toda evaluación clínica (Faust y Ziskin, 1989; Lanyon, 1984; Matarazzo, 1986, 1990; Tallent, 1987). Las principales preocupaciones incluyen las siguientes:

- Las pruebas psicológicas por computadora constituyen un sustituto pobre para la evaluación psicológica.
- Es poco común que los informes narrativos por computadora se validen antes de utilizarse.
- Las interpretaciones clínicas psicológicas por computadora no van firmadas.

Más adelante se matizan estos puntos.

Matarazzo (1986) ha enfatizado que la valoración por computadora es tan fácil de lograr y tiene una apariencia tan impresionante que tanto los usuarios como los receptores del informe narrativo pueden confundirla con una evaluación completa. Pero existe una diferencia entre prueba y evaluación. El clínico experimentado sabe que las pruebas psicológicas —ya sea tradicionales o por computadora— constituyen sólo el primer paso en una evaluación completa. Cuando ésta se lleva a cabo, el clínico competente va más allá de los resultados de la prueba para integrar las conclusiones dentro de la situación vital y antecedentes psicológicos completos de la persona examinada. En contraste, es poco común que un informe narrativo por computadora haga referencia a información externa a la prueba, como son el propósito de la evaluación, el funcionamiento adaptativo reciente del cliente o las impresiones de la entrevista, que podrían contradecir en gran medida las inferencias basadas en la prueba acerca de la personalidad.

En su crítica sobre las pruebas por computadora, Lanyon (1984) ha enfatizado que la interpretación impresa carece de validez demostrada. De hecho, la mayor parte de las interpretaciones automatizadas de una prueba se basan en el saber clínico más que en la validación empírica —es decir, tienen una naturaleza clínica más que actuarial. Aunque los comercializadores pueden llevar a cabo estudios de satisfacción de sus clientes (p. ej., "¿Qué tan satisfecho se encuentra usted con la afirmación narrativa *X* acerca del cliente *Y*?"), estos análisis no sustituyen el cuidadoso escrutinio de la validez interpretativa. Lanyon (1984) también señala con alarma que los sistemas de interpretación automática de prueba que están disponibles aumentan de manera exponencial. Casi cualquier persona puede comprar algunos de ellos para instalarlos en su computadora perso-

nal dentro del hogar. Estos avances predicen un sombrío futuro:

> Existe peligro real de que los muchos servicios insatisfactorios expulsen a aquéllos pocos que son satisfactorios, dado que los profesionales consumidores son, en general, incapaces de discriminar entre ellos y están predispuestos a creer cualquier cosa que esté impresa. En particular, es angustiante que la falta de validez demostrada para los programas se haya vuelto en la actualidad la norma y parece no haber una garantía en contra del mayor avance de esta situación insostenible. Quizá ha llegado el momento de que se creen reglamentos federales para esta industria, los cuales serían necesarios para la protección del consumidor (Lanyon, 1984).

Matarazzo (1986) hace sonar una alarma similar cuando advierte que la profesión de la psicología debe regularse o se arriesga a la incursión del Estado dentro de la práctica de las pruebas psicológicas.

Otro problema con las pruebas por computadora tiene que ver con el hecho de que es poco común que los informes narrativos automatizados vayan firmados. Los informes que no llevan una firma hacen surgir horrendas cuestiones con respecto a la responsabilidad profesional y la culpabilidad legal. Supóngase que un cliente resulta dañado a causa de un informe impreciso hecho por computadora. ¿Quién tiene la culpa? ¿Quién es legalmente responsable? La culpa puede residir en el psicólogo que utilizó el programa de computadora, en la empresa lucrativa que vendió el programa o en el programador individual que incorporó la afirmación (o afirmaciones) ofensiva dentro del *software*. Pero, en general, la responsabilidad legal reside en el psicólogo individual. La lección es que utilizar un sistema automatizado no absuelve al profesional de la responsabilidad sobre las consecuencias de un informe por computadora (ejemplo de caso 15–1).

A mediados del decenio de 1980, la *American Psychological Association* adoptó las *Guidelines for Computer-Based Tests and Interpretations* (1986). Estas pautas se reproducen en Butcher (1987, Apéndice B). El propósito general de las pautas es interpretar los *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1985) en su relación con las pruebas aplicadas e interpretadas por computadora. Dos pautas importantes incluyen lo siguiente:

1. Debe establecerse el grado en que las afirmaciones en un informe interpretativo se basan en investigación cuantitativa *versus* la opinión clínica experta.
2. Cuando las afirmaciones dentro de un informe interpretativo se basan en la opinión clínica experta, los usuarios deben recibir información que les permita ponderar la credibilidad de dicha opinión (Butcher, 1987).

Las pautas también detallan los muchos otros requisitos que deben satisfacer los autores, editores y usuarios de las pruebas. Quizás estos pasos hacia la autorregulación ayudarán a aumentar la respetabilidad y aceptabilidad de la redacción de informes por computadora. De otra manera, parecería inevitable un aumento en la incursión del Estado dentro de la práctica de las pruebas psicológicas.

PRUEBAS ADAPTATIVAS COMPUTARIZADAS

Una ventaja final de las pruebas por computadora es su aplicación a un examen flexible y adaptable. Las pruebas adaptativas no son nada nuevo —Binet las utilizó cuando desarrolló los métodos para encontrar los reactivos basales y superiores en su famosa prueba de inteligencia. Este autor colocó sus reactivos en un continuo de dificultad, de modo que el examinador pudiera someter a prueba en sentido inverso para encontrar el nivel basal de la persona evaluada y de manera ascendente para encontrar el nivel superior. Este procedimiento eliminó la necesidad de aplicar reactivos irrelevantes, aquéllos tan fáciles (por debajo del nivel basal) que la persona aprobaría con toda seguridad o aquéllos tan difíciles (por encima del nivel superior), que la persona fallaría con toda certeza. Otro ejemplo de prueba adaptativa es el procedimiento en dos etapas, donde se utilizan

EJEMPLO DE CASO 15–1
CONFIANZA IRREFLEXIVA EN UN INFORME DEL MMPI POR COMPUTADORA

En un caso civil, un mecánico demandó a un psicólogo clínico debido a que había confiado de manera inapropiada en una interpretación de prueba por computadora. El mecánico trabajaba para una gran compañía que poseía una flotilla privada de helicópteros. El trabajo del mecánico consistía principalmente en el mantenimiento y verificación de la seguridad de los helicópteros. Como parte de una política corporativa progresiva, el mecánico consultó con un psicólogo contratado por la compañía acerca de un problema relativamente menor que implicaba un conflicto familiar. El psicólogo le aplicó el MMPI al mecánico y envió el protocolo a un notable servicio de interpretación del MMPI. Pocos días después, llegó por correo una interpretación por computadora. Este informe estaba lleno de un texto narrativo alarmante. Mencionaba la enorme probabilidad de problemas con drogas y alcohol, la notable posibilidad de arranques explosivos y la fuerte probabilidad de pensamiento paranoide explícito. Se caracterizaba al cliente como una persona inestable, vengativa, paranoide y, posiblemente, peligrosa para sí misma y para los demás. Como es comprensible, el psicólogo se sintió alarmado. Afirmando tener el permiso del cliente, éste le comunicó a los directivos de la compañía los resultados del informe del MMPI. El supervisor del mecánico demandó que se realizaran valoraciones adicionales para abuso de alcohol y trastorno psicológico. Cuando el mecánico se negó, se le dio un permiso por enfermedad y finalmente se le despidió. En ese momento presentó una demanda en contra del psicólogo y de la compañía. El caso se arregló fuera del tribunal mediante una fuerte suma de dinero.

Lo que hace que este caso sea pertinente es que el informe por computadora discrepaba en gran medida con respecto al funcionamiento documentado del cliente. Aunque el mecánico reconocía un problema anterior con la bebida, ya no utilizaba drogas o alcohol. En su historial de trabajo no se encontró evidencia de arranques explosivos. Sus evaluaciones laborales anuales eran siempre satisfactorias. Sus supervisores y compañeros de trabajo lo describieron como carente de sentido del humor, pero no lo consideraban inestable, vengativo, paranoide o peligroso en ningún sentido. En apariencia, el psicólogo confió demasiado en la validez de la interpretación por computadora. El informe presentaba un tono excesivamente patológico. El error esencial del psicólogo consistió en aceptar de manera acrítica el informe, en vez de verificar su validez contra otras fuentes de información acerca del mecánico. Esto nos recuerda la conclusión cínica pero precisa de Lanyon (1984) acerca de los informes por computadora (señalada antes) en cuanto a que "los profesionales consumidores están en general... predispuestos a creer cualquier cosa que esté impresa". También se debe mencionar que no se consideró responsable al servicio de interpretación del MMPI. En última instancia, era el profesional quien tenía la responsabilidad de las consecuencias de la interpretación de prueba por computadora.

los resultados de una prueba directiva para determinar el nivel de entrada en escalas subsecuentes. Por ejemplo, en la Stanford-Binet: Cuarta Edición, los resultados de la subprueba inicial de vocabulario determinan el punto de inicio para las subpruebas subsecuentes. Al reducir el tiempo necesario para obtener una medida precisa de la capacidad, las pruebas adaptativas satisfacen un propósito muy constructivo.

Las **pruebas adaptativas computarizadas** (CAT; del inglés *computarized adaptative testing*) constituyen una familia de procedimientos que permiten la medición precisa y eficaz de la capacidad (Weiss, 1985; Weiss y Vale, 1987). Aunque los detalles difieren de un método a otro, la mayor parte de las formas de las pruebas adaptativas computarizadas comparten las siguientes características:

1. Con base en un extenso método de preprueba, se valoran de manera precisa las características de respuesta de cada reactivo (porcentaje de aprobación *versus* capacidad).

2. Estas características de respuesta de los reactivos se programan en la computadora junto con una estrategia CAT de selección de reactivos.
3. En la selección del siguiente reactivo que se presenta, la computadora utiliza la historia total de respuestas realizadas hasta ese momento por la persona examinada.
4. Después de cada respuesta, la computadora vuelve a calcular el nivel estimado de capacidad de la persona examinada.
5. La computadora también estima la precisión de la medición (p. ej., error estándar de medición) después de cada respuesta.
6. La prueba continúa hasta que se alcanza un nivel predeterminado de precisión en la medición.
7. La calificación de la persona examinada se basa en el nivel de dificultad y en otras características de medición de los reactivos aprobados, no en el número total de aquéllos correctos.

Las ventajas de medición de las CAT pueden resumirse en dos palabras: precisión y eficiencia (Weiss y Vale, 1987). En cuanto a la precisión, las CAT garantizan que se mida a cada persona con el mismo grado de precisión debido a que la prueba continúa hasta que este criterio se satisface. Esto no ocurre con las pruebas tradicionales, donde las calificaciones en ambos extremos de la distribución reflejan mayores niveles de error de medición que aquéllas al centro de la misma. En cuanto a eficiencia, el enfoque de las CAT requiere de muchos menos reactivos de los que se necesitan en las aplicaciones tradicionales. Por ejemplo, los exámenes escritos para certificación incluyen por lo común de 200 a 500 reactivos, mientras que los CAT son siempre más cortos y con frecuencia incluyen menos de 100 reactivos para lograr un nivel más preciso de medición (Lunz y Bergstrom, 1994). En un análisis, la confiabilidad de las pruebas adaptativas computarizadas alternativas para la certificación en tecnología médica fue de .96 (Lunz, Bergstrom y Wright, 1994). Esto es notable, ya que las pruebas más breves (que es la meta en las CAT) tienden a ser menos confiables que aquéllas más extensas (como las que se encuentran en los programas tradicionales de prueba).

El enfoque de las CAT para el examen psicológico se ha utilizado principalmente en grandes instituciones como el Ejército de EUA y el *Educational Testing Service*, para la evaluación de la inteligencia y capacidades especiales. En años recientes, las juntas nacionales de licencias (p. ej., en medicina) han comenzado a poner en práctica los exámenes CAT debido a la conveniencia en la programación de pruebas, mayor control sobre la seguridad de los exámenes, reducción en costos y oportunidad de obtención de mejores datos (Lunz y Bergstrom, 1994). La información técnica sobre los sistemas CAT está patentada y es difícil de obtener. Sin embargo, es claro que la eficiencia del enfoque CAT es sustancial. Éstas utilizan menos reactivos de mejor calidad que una prueba convencional de la misma extensión. Una conclusión general es que las CAT reducen la extensión de una prueba en cerca de 50%, con disminuciones para los individuos examinados de hasta 80%, sin pérdida de la precisión en la medida (Laatsch y Choca, 1994; Weiss y Vale, 1987).

Del mismo modo en conforme que los costos de las computadoras continúan disminuyendo de manera notable, se desarrollarán aplicaciones a mayor escala de las CAT. A finales del decenio de 1990, el *Educational Testing Service* avanzó hacia una confiabilidad casi total en las versiones CAT del *Graduate Record Examination* (Examen de Registro de Posgrado) y otras pruebas de selección. Las juntas de licencias y certificación como el *National Council of State Boards of Nursing* también han introducido versiones CAT en sus pruebas de certificación. Mills y Stocking (1996) analizan las cuestiones prácticas de las pruebas computarizadas aplicadas a gran escala.

FUTURO DE LA APLICACIÓN DE PRUEBAS

En los inicios del siglo XXI, ¿cuál es el futuro de la aplicación de pruebas psicológicas? A

continuación se arriesgan algunas especulaciones, a pesar de saber que con frecuencia los pronósticos del futuro son incorrectos. La predicción de los avances en pruebas es especialmente difícil debido a que la opinión pública restringe cada vez más, ya sea de manera directa o indirecta, esta empresa. Por ejemplo, en algún momento durante el decenio de 1980, la legislatura del estado de California determinó que era **ilegal** que los psicólogos escolares utilizaran pruebas tradicionales de inteligencia como base para asignar a los alumnos a clases de educación especial. Estas restricciones sobre las pruebas fueron el producto de la indignación pública acerca de la asignación excesiva de estudiantes de minorías a clases de educación especial. Así, aunque sea factible una tecnología particular de las pruebas y la promuevan los psicólogos, existe la posibilidad de que se le controle de manera estricta o, incluso, que se le prohíba.

Un caso específico a propósito del tema se encuentra en la predicción de Matarazzo (1992) acerca de que las medidas biológicas de la inteligencia obtendrán un lugar más destacado en el siglo XXI. Con seguridad, parece cierto que las medidas biológicas de la capacidad, como el potencial evocado promedio (estimado a partir de las ondas en el EEG) o la tasa metabólica de la glucosa en el cerebro (estimada a partir de rastreos TEP) o el tamaño relativo del cerebro (estimado a partir de rastreos de IRM), resultarán enfoques efectivos para la evaluación (véase el tema 5A, Teorías y la medición de la inteligencia). Pero Matarazzo (1992) va un paso más allá al afirmar que, de hecho, éstos y otros enfoques biológicos se volverán de uso común:

> Por tanto, otra de mis predicciones es que en los primeros decenios del siglo XXI es posible que veamos un mayor desarrollo y **utilización dentro de la práctica** de éstos y otros índices del funcionamiento y estructura cerebral en una prueba (o batería de pruebas) para la medición de las diferencias individuales en capacidad mental, con lo cual se anunciará la primera separación clara en un siglo con respecto a los reactivos de una prueba y exámenes de la tradición de Binet (p. 1012, negritas tomadas del original).

Aunque el pronóstico de Matarazzo puede volverse realidad, un escenario más probable es que el público en general se sienta amenazado cuando se utilicen los índices biológicos en la evaluación y, por tanto, tomará medidas (p. ej., presionar a los legisladores) para garantizar que dichas medidas reciban aplicación limitada (o nula). El público se sentirá amenazado debido a que, de manera correcta o incorrecta, las características biológicas como la tasa metabólica de la glucosa en el cerebro se perciben como relativamente permanentes e inmutables. Surgirá el temor de que las pruebas biológicas clasifiquen a las personas en sistemas de castas. Incluso si se establece (o cuando se establezca) de manera firme la validez de las pruebas biológicas, pasarán decenios (si es que acaso algún día ocurre) antes de que el público general las encuentre aceptables.

Por otro lado, la conversión de pruebas a versiones por computadora es una parte integral de las sociedades industrializadas y esta tendencia aumentará en el futuro. Las pruebas existentes se adaptarán a las computadoras personales con una regularidad cada vez mayor. Un ejemplo de esta tendencia es la Fepsy (*Ferrum* + *Psyche*), un sistema de pruebas neuropsicológicas automatizadas que se encuentra disponible en línea en 220 sitios a través de Holanda y la mayor parte de Europa. La descripción del Fepsy se encuentra en la *internet* en la dirección www.euronet.nl Fepsy consiste de las siguientes subpruebas:

- Tiempo de reacción auditiva.
- Tiempo de reacción para elección binaria.
- Tarea de golpeteo.
- Tarea de búsqueda visual.
- Tareas de reconocimiento.
- Tarea de vigilancia.
- Tarea rítmica.
- Tarea de clasificación.
- Tareas de medios campos visuales.
- Golpeteo con cubos de Corsi.

Un empleo común es en la evaluación pre y posoperatoria de pacientes que atraviesan por

cirugía para la epilepsia, con el propósito de aliviar las convulsiones. El sistema se ha utilizado con pacientes completamente conscientes **durante** la cirugía. Bajo anestesia local, el paciente trabaja en una subprueba mientras que, de manera simultánea, recibe estimulación eléctrica inofensiva en sitios distintivos de la corteza. El propósito consiste en determinar si la ablación de tejido cicatricial del cerebro podría afectar funciones cognitivas específicas. La ventaja de utilizar un sistema multicentral, multinacional, de pruebas por computadora es que el examinador tiene acceso a datos normativos de **miles** de pacientes con padecimientos específicos.

Otra predicción es que las editoriales de pruebas publicarán cada vez menos pruebas de amplio espectro (p. ej., inventarios de personalidad y pruebas individuales de inteligencia) (Gregory, 1998). En vez de ello, se concentrarán en pruebas diseñadas para evaluar áreas particulares de funcionamiento para poblaciones específicas (p. ej., medidas del funcionamiento de la memoria para ancianos de los que existe sospecha de demencia). Las razones para estas tendencias complementarias son de naturaleza económica:

> La publicación de pruebas es un gran negocio, una manera respetable de que las grandes corporaciones obtengan una ganancia. Las editoriales se mostrarán renuentes a hacer la enorme inversión que se necesita para desarrollar nuevos instrumentos que tienen la grandiosa ambición de evaluar muchos aspectos de la personalidad o intelecto en una amplia gama de personas. El costo es demasiado elevado y —ante la competencia existente— el riesgo es demasiado grande [Gregory, 1998, pp. 76-77].

Es probable que los editores de pruebas se enfoquen en formas menos costosas y arriesgadas de desarrollo de pruebas, como en instrumentos que representan constructos distintivos relevantes para grupos específicos. Los ejemplos podrían incluir pruebas para medir conductas de riesgo en adolescentes, disminución mental en ancianos, cogniciones defectuosas en personas con depresión, o problemas de comunicación en parejas con dificultades maritales. Estos tipos de instrumentos florecerán, mientras que los editores raras veces invertirán en nuevas pruebas generales de personalidad o capacidad, prefiriendo en su lugar la revisión y reciclaje de los instrumentos existentes.

RESUMEN

1. El término **evaluación psicológica asistida por computadora** (CAPA) se refiere al rango completo de aplicaciones computacionales en la evaluación psicológica. Éstas incluyen la aplicación, calificación e interpretación de pruebas; pruebas adaptativas computarizadas y aplicaciones refinadas en multimedia.
2. El primer uso de las computadoras para proporcionar interpretaciones se puede encontrar en la Clínica Mayo a principios del decenio de 1960. El sistema interpretativo del MMPI proporcionaba afirmaciones breves, escala por escala, basadas en el conocimiento clínico.
3. En la actualidad se encuentran disponibles interpretaciones de prueba basadas en sistemas (CBTI) para casi cualquier prueba psicológica publicada. Se reconocen cuatro enfoques para ésta: informes de calificación, informes descriptivos, informes actuariales e informes clínicos asistidos por computadora.
4. Los informes de calificación consisten sólo de las calificaciones, perfiles o ambos, pero pueden incluir la representación gráfica de pruebas de significación estadística e intervalos de confianza para las calificaciones de una prueba. Estos informes destacan a primera vista las calificaciones significativas y las diferencias en puntuación.
5. Un informe descriptivo proporciona una interpretación breve, escala por escala, de los resultados de la prueba. Éstos son espe-

cialmente útiles cuando las conclusiones de una prueba se trasmiten a profesionales de la salud mental que tienen poco conocimiento de la prueba en cuestión.

6. En la interpretación actuarial de una prueba se utiliza una fórmula derivada a nivel empírico para diagnosticar, clasificar o predecir la conducta. Este tipo de interpretación contrasta con el enfoque clínico, en el que el psicólogo procesa la información dentro de su cabeza para diagnosticar, clasificar o predecir la conducta.
7. Las comparaciones empíricas entre la interpretación clínica y la intepretación actuarial de pruebas han encontrado que esta última es superior en casi todos los casos. Las interpretaciones de una prueba por computadora deberían incorporar los métodos actuariales siempre que sea posible.
8. En un informe clínico asistido por computadora, las afirmaciones interpretativas se basan en el juicio automatizado y computarizado de uno o más expertos clínicos. Este enfoque permite la interpretación de todos los perfiles de una prueba, no sólo de aquellos que encajan con ciertos patrones actuariales.
9. Las ventajas de la interpretación de una prueba basada en sistemas incluye la objetividad, velocidad y bajo costo. Una de las principales desventajas es el peligro de que al psicólogo se le excluya por completo del proceso de evaluación, lo cual aumenta el riesgo del uso indebido de los resultados de una prueba.
10. Los sistemas multimedia incluyen la presentación realista, interactiva, de los estímulos de una prueba por computadora (p. ej., representación en vídeo de una situación de trabajo). Multimedia permite que se someta a prueba a los individuos en cuanto a problemas complejos, de la vida real, como la solución de conflictos en el trabajo.
11. Las pruebas adaptativas computarizadas (CAT) son una familia de procedimientos que permite la medición precisa y eficaz de la capacidad. En este enfoque, la computadora guía la selección de reactivos con base en las respuestas anteriores de la persona examinada.
12. El propósito de las CAT consiste en llegar a un nivel predeterminado de precisión de la medición con el menor número posible de reactivos de una prueba. Una conclusión típica es que las CAT reducen la extensión de una prueba en cerca de 50%, sin pérdida de precisión en la medida.
13. Es difícil pronosticar el futuro de las pruebas. Mientras que algunas autoridades predicen un incremento en la utilización de medidas biológicas de la inteligencia, esto es incierto. Con seguridad, la automatización cada vez mayor de las pruebas es una tendencia clara que sólo puede intensificarse.

TÉRMINOS Y CONCEPTOS CLAVE

Evaluación psicológica asistida por computadora p. 664
Interpretación de prueba basada en sistemas p. 665
Juicio clínico p. 669
Juicio actuarial p. 669
Multimedia p. 674
Realidad virtual p. 675
Pruebas adaptativas computarizadas p. 679

TEMA 15B
Cuestiones éticas y sociales en las pruebas

Fundamentación para las normas profesionales de prueba
Ejemplo de caso 15–2. Dilemas éticos y profesionales en las pruebas
Responsabilidades de los editores de pruebas
Responsabilidades de los usuarios de pruebas
Ejemplo de caso 15–3. Entusiasmo extremo en la interpretación del MMPI
Pruebas con minorías culturales y lingüísticas
Reiteración: uso responsable de las pruebas
Resumen
Términos y conceptos clave

El principal punto de vista que se enfatiza en esta obra es que las pruebas psicológicas han tenido una influencia benéfica en la sociedad moderna. Cuando se les emplea de manera ética y responsable, las pruebas proporcionan una base para llegar a inferencias sensatas acerca de individuos y grupos. Después de todo, la intención de la tarea consiste en promover la orientación apropiada, tratamiento eficaz, valoración precisa y toma de decisiones equitativas —ya sea en las pruebas individuales o en las evaluaciones grupales institucionales. ¿Quién podría tener queja sobre estos fines?

Por fortuna, los psicólogos, educadores, administradores y otras personas en general aplican las pruebas de modo responsable. Pero existen excepciones. Casi todos hemos escuchado horribles anécdotas, el alumno de primaria proveniente de una minoría y que por casualidad ha recibido una etiqueta de retraso mental con base en una sola calificación de CI; el estudiante universitario con un diagnóstico poco plausible de esquizofrenia a partir de una prueba proyectiva; el solicitante de un empleo al que se ha descartado erróneamente de un trabajo con base en una medida irrelevante; el as-

pirante a maestro que ha recibido una ventaja injusta cuando de antemano una prueba de competencia profesional se ha perdido de manera misteriosa; o el niño de una minoría a quien se ha penalizado en una prueba debido a que el inglés no es su lengua materna. Excepciones como éstas ilustran la necesidad de normas éticas y profesionales en la aplicación de pruebas.

Uno de los principales propósitos de este tema es introducir al lector a las normas éticas y profesionales que dan forma a la práctica de las pruebas psicológicas. También se analiza el tema relacionado de las consideraciones especiales en la evaluación de minorías culturales y lingüísticas. Ambos temas comparten una superposición sustancial: cuando una persona evaluada no proviene de la cultura angloestadounidense mayoritaria (que predominantemente es caucásica, angloparlante, individualista, orientada al futuro), se elevan a un primer sitio las preocupaciones éticas y profesionales en la aplicación de pruebas.

FUNDAMENTACIÓN PARA LAS NORMAS PROFESIONALES DE PRUEBA

En general, las pruebas se aplican de manera responsable, pero como se señaló con anterioridad, existen excepciones. En pocas ocasiones, el proceso de prueba es irresponsable a causa del diseño más que por accidente. Considérese, con asombro estremecedor, el anuncio de la Sonda Mental presentado en una revista de psicología popular:

> ¿Últimamente ha leído alguna mente interesante? Con la Sonda Mental puede hacerlo. Con sólo unos minutos tendrá el perfil de personalidad de cualquiera con precisión científica. Este nuevo *software* de sistemas expertos le permite descubrir las cosas que la mayoría de la gente teme decirle. Las fortalezas, debilidades, intereses sexuales y más (Eyde y Primhoff, 1992).

En este caso, la irresponsabilidad es tan flagrante que discutir sobre pautas éticas y profesionales es casi superfluo.

No obstante, no siempre se presenta en agudos matices contrastantes, responsables o irresponsables. El verdadero reto de la evaluación competente consiste en determinar las fronteras de la práctica ética y profesional. Como suele suceder, son los casos que están en el límite los que dan lugar a consideraciones. Se alienta al lector para que lea los dilemas asociados con las pruebas que se describen en el Ejemplo de caso 15–2 y se forme una opinión acerca de cada uno. Estos ejemplos se basan en informes de primera mano del autor. Al final de este capítulo se regresará a estas anécdotas problemáticas.

Los dilemas de la examinación psicológica no siempre tienen respuestas simples y obvias. Incluso los psicólogos cuidadosos y experimentados pueden estar en desacuerdo acerca de lo que es ético y profesional en un caso específico. Sin embargo, el alcance de la práctica profesional y ética no es cuestión de gustos individuales o juicio personal. El empleo responsable de pruebas está definido por pautas escritas y publicadas por asociaciones profesionales como la *American Psychological Association*, la *American Counseling Association*, la *National Association of School Psychologists* y otros grupos. Lo sepan o no, todos los profesionales practicantes deben lealtad a estas pautas, las cuales se revisan más adelante.

En general, la evolución de las normas éticas y profesionales ha sido restrictiva casi de manera uniforme, al proporcionar una delimitación cada vez más estrecha de dónde, cuándo y cómo pueden utilizarse las pruebas psicológicas. Desde una perspectiva legal, Bersoff (1984) resumió las tendencias históricas como sigue:

> En algún momento, el trabajo de los psicómetras académicos y aplicados pasaba casi sin una revisión por parte de la ley, pero a medida que el uso de pruebas aumentó en EUA, también se incrementó su potencial para causar un daño a las personas examinadas que era competencia de la ley. Como resultado, es probable que en la actualidad no exista una actividad realizada por los psicólogos que esté bajo un escrutinio tan ceñido y bajo tal reglamentación del sistema legal como lo son las pruebas.

EJEMPLO DE CASO 15–2
DILEMAS ÉTICOS Y PROFESIONALES EN LAS PRUEBAS

1. Un psicólogo consultor está de acuerdo en llevar a cabo un examen de psicopatología como selección previa al empleo de candidatos a oficiales de policía. Al principio de cada consulta, el psicólogo le pide al candidato que lea y firme una forma de consentimiento que describe de manera abierta y sincera el proceso de evaluación. No obstante, dicha forma explica que no se le proporcionará a los candidatos al empleo la información específica acerca de los resultados de la prueba. Pregunta: ¿es ético que el psicólogo niegue a los candidatos tal información de resultados?
2. Un orientador competente que ha recibido amplio entrenamiento en la interpretación del MMPI continúa utilizando este instrumento aunque lo ha sustituido el MMPI-2. Su fundamentación es simplemente que existe gran cantidad de investigación sobre el MMPI y él tiene confianza en el significado de las elevaciones en los perfiles de esta prueba, mientras que sabe muy poco acerca del MMPI-2. Pregunta: ¿la negativa del orientador a utilizar el MMPI-2 quebranta las normas profesionales?
3. Se pide a un psicólogo consultor que valore a un niño de nueve años de edad, de origen puertorriqueño, para examinar una posible discapacidad para el aprendizaje. El idioma principal del niño es el español, y su segunda lengua es el inglés. El psicólogo intenta utilizar la Escala Wechsler de Inteligencia para Niños–III (WISC-III) y otras pruebas. Debido a que no sabe español, el psicólogo le pide a quien cuida del niño cuando éste regresa de la escuela, que actúe como traductora cuando se necesiten comunicar instrucciones de la prueba, preguntas específicas o las respuestas del niño. Pregunta: ¿es una práctica apropiada utilizar a un traductor cuando se aplica una prueba individual como el WISC-III?
4. Cuando respondía la mitad de una batería de pruebas para discapacidad para el aprendizaje, una angustiada estudiante universitaria de 20 años de edad le confiesa un terrible secreto al psicólogo. La cliente acaba de descubrir que es muy probable que su hermano de 25 años, quién murió tres meses antes, haya sido pedófilo. La estudiante le muestra al psicólogo fotografías de niños desnudos posando en la habitación de su hermano. Para complicar las cosas, el hermano vivía con su madre —quien aún desconoce su bien ocultada desviación sexual. Pregunta ¿el psicólogo está obligado a informar de este caso a las autoridades judiciales?

En parte como respuesta al clima actual de querellas legales, las organizaciones que se ocupan de las pruebas psicológicas han publicado pautas que de manera colectiva definen las normas éticas y profesionales relativas a la práctica de la evaluación.

Estas normas también se relacionan con las corporaciones e individuos que publican pruebas. Se comienza con una exploración de las pautas para los editores de pruebas antes de analizar las responsabilidades de los usuarios de las mismas. El capítulo se cierra con una revisión de las preocupaciones especiales en la evaluación de minorías culturales y lingüísticas.

RESPONSABILIDADES DE LOS EDITORES DE PRUEBAS

Las responsabilidades de los editores tienen que ver con la publicación, mercadeo y distribución de sus pruebas. En particular, se espera que publiquen pruebas de alta calidad, mercadeen

su producto de manera responsable y restrinjan su distribución sólo a personas con los requisitos adecuados. Se considerarán cada uno de estos temas.

CUESTIONES DE PUBLICACIÓN Y MERCADEO

En cuanto a la publicación de instrumentos nuevos o revisados, la pauta más importante consiste en protegerse contra la difusión prematura de una prueba. Éstas son una empresa noble, pero también son un gran negocio impulsado por las ganancias, lo cual resulta en una presión inherente hacia la pronta difusión de materiales nuevos y revisados. Quizás ésta es una de las razones por las que la *American Psychological Association* y otras organizaciones han publicado normas que se relacionan con la publicación de pruebas (AERA/APA/NCME, 1985, 1999). Dichas normas se relacionan en particular con los manuales técnicos y guías para usuarios que por lo común acompañan a una prueba. Estas fuentes deben ser suficientemente completas, de modo que un usuario o revisor calificados puedan valorar la pertinencia e idoneidad técnica de la prueba. Esto significa que los manuales y guías informarán en detalle los datos estadísticos sobre análisis de confiabilidad, estudios de validez, muestras normativas y otros aspectos técnicos.

El mercadeo de pruebas de manera responsable se refiere no sólo a la publicidad (que debe ser precisa y formal), sino también al modo en que se presenta la información en los manuales y guías. En particular, los autores de pruebas deben esforzarse en lograr una presentación equilibrada de sus instrumentos y abstenerse de una manifestación unilateral de la información. Por ejemplo, si algunos estudios preliminares reflejan una imagen deficiente de una prueba, se les debe dar una ponderación justa en el manual junto con las conclusiones positivas. De la misma manera, si se puede anticipar el empleo incorrecto o inapropiado de una prueba, es necesario que el autor analice también esta cuestión.

COMPETENCIA DE LOS COMPRADORES DE PRUEBAS

Los editores de pruebas reconocen la enorme responsabilidad de que sólo los usuarios calificados deben tener la posibilidad de comprar sus productos. A manera de una breve revisión (véase el tema 2A, Naturaleza y usos de las pruebas psicológicas), las razones para el acceso restringido incluyen el potencial de daño si las pruebas caen en manos inadecuadas (p. ej., un estudiante de pregrado en psicología aplica el MMPI-2 a sus amigos y después hace pronunciamientos aterradores acerca de los resultados) y el hecho obvio de que muchas de ellas ya no pueden considerarse válidas si el examinando potencial las ha revisado de antemano (p. ej., un maestro memoriza las respuestas correctas para un examen de certificación).

Estos ejemplos ilustran que el acceso a las pruebas psicológicas debe ser limitado. Pero, ¿limitado para quiénes? Resulta ser que la respuesta depende de la complejidad de la prueba específica bajo consideración. Las pautas propuestas hace muchos años por la *American Psychological Association* (APA) continúan siendo pertinentes hoy en día, a pesar de que no todos los editores las ponen en práctica. La APA propone que las pruebas se clasifiquen en tres niveles de complejidad (A, B y C) que requieren diferentes grados de conocimiento experto por parte del examinador.

- **Nivel A:** estos instrumentos son medidas sencillas de lápiz y papel que pueden aplicarse, calificarse e interpretarse con un mínimo de entrenamiento. Con ayuda de un manual, personas responsables sin entrenamiento en psicología, como ejecutivos comerciales o administradores educativos, pueden utilizar estas pruebas. Esta categoría incluye pruebas grupales de destreza vocacional y de aprovechamiento educativo.
- **Nivel B:** éstas requieren conocimiento sobre construcción de pruebas y entrenamiento en estadística y psicología. Estos productos están disponibles para personas que

han terminado cursos avanzados en pruebas en una escuela o universidad acreditada, o entrenamiento equivalente bajo la supervisión de un psicólogo calificado. Esta categoría incluye pruebas de aptitud e inventarios de personalidad aplicables a poblaciones normales.

Nivel C: estas pruebas requieren de comprensión sustancial sobre las pruebas y temas asociados. La experiencia bajo supervisión es esencial para la aplicación, calificación e interpretación apropiadas de estos instrumentos. Por lo común, las pruebas del Nivel C están disponibles sólo para personas con un mínimo de maestría en psicología o en un campo relacionado. Estos instrumentos incluyen pruebas individuales de inteligencia, pruebas proyectivas de la personalidad y baterías de pruebas neuropsicológicas (*American Psychological Association*, 1953).

En general, los editores de prueba tratan de descartar las solicitudes inapropiadas al solicitar a los compradores las credenciales necesarias. Por ejemplo, la *Psychological Corporation*, uno de los principales proveedores de materiales de prueba en EUA, requiere que los futuros consumidores llenen una forma de registro en la que detallen su entrenamiento y experiencia con pruebas. Los compradores que no posean un título avanzado en psicología deben listar en detalle los cursos sobre aplicación e interpretación de pruebas y sobre estadística. También se solicitan referencias.

La mayoría de los editores de pruebas también especifican que los individuos o grupos que proporcionan pruebas y orientación por correo no pueden adquirir los materiales. En una nota relacionada, las normas éticas desalientan en la actualidad la aplicación de pruebas "para llevar a casa" con sus clientes. Hasta hace pocos años, ésta era una práctica ocasional con las pruebas extensas de personalidad como el MMPI. El comité de ética refrendó los siguientes puntos:

1. En general, la aplicación sin vigilancia del MMPI no representa una práctica responsable de prueba y puede resultar en una evaluación inválida por diversas razones (p. ej., influencia de otras personas o responder la prueba en un estado de intoxicación).
2. La seguridad de la prueba se pone en entredicho cuando se permite que el MMPI salga del ambiente clínico.
3. Existe debate en cuanto a si hay alguna circunstancia en la que podría considerarse razonable y apropiado permitir que el MMPI se responda fuera del ambiente clínico.
4. Estas cuestiones no son únicas del MMPI, sino que deben considerarse cuando se realiza cualquier evaluación.
5. Cuando se juzgan las cuestiones éticas asociadas con la aplicación de pruebas en el hogar, es importante considerar cosas tales como la naturaleza y propósito de la prueba y la información disponible sobre confiabilidad, validez y procedimientos de estandarización (APA, 1994b, pp. 665-666).

En general, se aconseja que los usuarios no apliquen pruebas "para llevar a casa" y se exhorta a los editores a negar acceso a los profesionales o grupos que promuevan esta práctica.

Aunque los editores intentan filtrar a los compradores no calificados, de todas maneras pueden existir casos en los que se venden pruebas confidenciales a individuos poco escrupulosos. Oles y Davis (1977) descubrieron que los estudiantes de posgrado en psicología podían comprar el WISC-R, MMPI, TAT, Stanford-Binet y 16FP si escribían las órdenes de compra en la papelería de la universidad, colocaban las letras *Ph.D.* después de sus nombres, incluían el pago y utilizaban una dirección en una oficina postal. Aunque las órdenes ilícitas de prueba son pocas en cuanto a número, sí llegan a ocurrir.

RESPONSABILIDADES DE LOS USUARIOS DE PRUEBAS

La evaluación psicológica de la personalidad, intereses, funcionamiento cerebral, aptitudes o

inteligencia constituye un acto profesional delicado que debería realizarse con supremo cuidado del bienestar de la persona examinada, su familia, empleadores y la red más amplia de instituciones sociales que podrían resultar afectadas por las conclusiones de una evaluación clínica particular (Matarazzo, 1986, 1990). A través de los años, la profesión de la psicología ha propuesto, aclarado y estrechado una serie de normas acabadas y cuidadosas para proporcionar una guía al practicante individual. Las organizaciones profesionales publican principios éticos formales que conciernen al uso de pruebas, incluyendo la *American Psychological Association* (APA, 1992), la *American Association for Counseling and Development* (AACD, 1988), la *American Speech Language-Hearing Association* (ASHA, 1991) y la *National Association of School Psychologists* (NASP, 1992).

Además de los principios éticos, varias organizaciones relacionadas con el uso de pruebas han publicado normas prácticas que ayudan a definir el espectro del empleo responsable de pruebas. Las fuentes de las pautas para su uso incluyen grupos de enseñanza (AFT, NCME, NEA, 1990), la *American Psychological Association* (APA, 1992b), el *Education Test Service* (ETS, 1987, 1988, 1989), el *Joint Committee on Testing Practices* (JCTP, 1988), la *Society for Industrial and Organizational Psychology* (SIOP, 1987) y las asociaciones profesionales (AERA, APA y NCME, 1985, 1999). Por último, debería mencionarse que los principios del uso responsable de las pruebas se han sintetizado en un ilustrativo registro de casos publicado de manera conjunta por varios grupos relacionados con la aplicación de pruebas (Eyde, Robertson, Krug y otros, 1993).

Las docenas de pautas relativas al uso de pruebas son bastante específicas, por ejemplo:

> Norma 6.8: cuando los resultados de una prueba se proporcionan a medios noticiosos, aquéllos responsables de la entrega de los resultados deben brindar información que ayude a reducir al mínimo la posibilidad de interpretaciones erróneas de dichos resultados (AERA, APA Y NCME, 1985).

Debido a su especificidad, un análisis detallado de las normas éticas y profesionales pertinentes está más allá del campo de acción de este libro. Lo que sigue es un resumen de las disposiciones generales que se relacionan con la práctica responsable de la aplicación de pruebas psicológicas y de la evaluación en psicología clínica.

Estos principios se aplican a los psicólogos, estudiantes de psicología y otras personas que trabajan bajo la supervisión de un psicólogo. Esta discusión se restringe aquellos principios que tienen relación directa con la práctica de las pruebas psicológicas. La observancia adecuada de estos preceptos puede eliminar la mayor parte —si no es que todos— los desafíos legales del uso de pruebas.

PROTECCIÓN DE LOS MEJORES INTERESES DEL CLIENTE

Varios principios éticos reconocen que todos los servicios psicológicos, incluyendo la evaluación, se proporcionan dentro del contexto de la relación profesional. Por tanto, los psicólogos están obligados a aceptar la responsabilidad implícita en esta relación. En general, el profesional se guía por una pregunta preponderante: ¿cuáles son los mejores intereses del cliente? La implicación funcional de esta pauta es que la evaluación debe satisfacer un propósito constructivo para la persona examinada. De no ser así, es probable que el profesional esté violando uno o más principios éticos específicos. Por ejemplo, la Norma 6.5 del manual de **Normas** (AERA, APA Y NCME, 1985) advierte a los usuarios de pruebas que deben evitar acciones que puedan tener consecuencias negativas no intencionales. Permitir que un cliente dé significados adicionales no sustentados a los resultados de una prueba iría en contra de los intereses del cliente y, por tanto, constituiría una práctica poco ética de aplicación de pruebas. De hecho, con ciertos clientes propensos a la preocupación y a tener dudas acerca de ellos mismos, un psicólogo podría elegir no utilizar una prueba apropiada, dado que es casi seguro

que estos clientes interpreten de manera errónea y autodestructiva casi cualquier resultado de prueba.

CONFIDENCIALIDAD Y OBLIGACIÓN DE ADVERTIR

Los profesionales tienen la obligación primordial de proteger la confidencialidad de la información, incluyendo los resultados de una prueba, que puedan obtener de sus clientes en el curso de una consulta (Principio 5, APA, 1992a). Dicha información puede comunicarse de manera ética a otras personas sólo después de que el cliente o su representante legal dan su consentimiento sin ambigüedades, en general por escrito. Las únicas excepciones a la confidencialidad implican aquellas circunstancias poco comunes en las que retener información podría representar un peligro evidente para el cliente o para otras personas. Por ejemplo, la mayor parte de los estados en EUA han aprobado leyes que obligan a los profesionales de la salud a informar de todos los casos en los que se sospeche la existencia de maltrato a niños y de ancianos bajo circunstancias vulnerables. En la mayor parte de los estados, un psicólogo que se entera en el curso de la aplicación de pruebas que el cliente ha abusado física o sexualmente de un niño, tiene la obligación de dar esta información a la víctima potencial, a los agentes de la ley, o a ambos.

Los psicólogos también tienen una **obligación de advertir** que surge de la decisión de 1976 para el caso Tarasoff (Wrightsman, Nietzel y Fortune, 1994). Tanya Tarasoff, una joven estudiante universitaria en California asesinada por Prosenjit Poddar, un estudiante de la India. Lo que hace que este caso sea pertinente para la práctica de la psicología es que Poddar le había manifestado al terapeuta de la universidad sus intenciones de asesinar a Tarasoff. Aunque el terapeuta le advirtió a la policía que Poddar había expresado estas amenazas, no le avisó a Tarasoff. Dos meses después, Poddar apuñaló y mató a la joven en el hogar de ésta. Los padres de Tarasoff presentaron una demanda y la Suprema Corte de California coincidió en que los terapeutas tienen la obligación de utilizar "precaución razonable" para proteger a las potenciales víctimas de sus clientes. Aunque la resolución Tarasoff ha sufrido modificaciones legislativas en muchos estados, el impulso del caso aún se mantiene: los clínicos deben comunicar cualquier amenaza seria tanto a la víctima potencial como a las autoridades competentes, o a ambos.

Por último, el clínico debe considerar el bienestar del cliente cuando decide liberar información, en especial cuando éste es un menor que no puede dar su consentimiento voluntario e informado. Cuando sea apropiado, los profesionales deben informar a sus clientes de los límites legales de la confidencialidad.

DESTREZA DEL USUARIO DE PRUEBAS

Varios principios reconocen que el usuario de una prueba debe aceptar la responsabilidad última de la aplicación apropiada de las mismas. Desde un punto de vista práctico, esto significa que la persona que emplea una prueba debe tener el entrenamiento adecuado en evaluación y teoría de medición. El usuario debe poseer la destreza necesaria para valorar las pruebas psicológicas en cuanto a la exactitud de su estandarización, confiabilidad, validez, precisión interpretativa y otras características psicométricas. Esta norma es de particular importancia en áreas como la selección laboral, educación especial, examen de personas con discapacidades u otras situaciones donde el impacto potencial es enorme.

Los psicólogos que tienen un entrenamiento deficiente en los instrumentos elegidos, pueden cometer graves errores de interpretación que dañen a las personas examinadas. Lo que es más, la utilización inepta de pruebas puede exponer al examinador a sanciones profesionales y demandas de tipo civil. Un error común que se observa entre los usuarios sin experiencia es la interpretación demasiado entusiasta, patologizante, de los resultados de pruebas de personalidad (ejemplo de caso 15–3).

EJEMPLO DE CASO 15–3. ENTUSIASMO EXTREMO EN LA INTERPRETACIÓN DEL MMPI

Un psicólogo consultor sin experiencia utilizaba de manera rutinaria el MMPI con propósitos de selección previa a la contratación de los candidatos a agentes policiacos. Una candidata presentó una demanda subsecuentemente, alegando daños a causa del informe del psicólogo. A la demandante, una mujer joven con amplio entrenamiento y antecedentes dentro del sistema de imposición de la ley, se le negó un puesto como oficial de policía debido a que supuestamente tenía un perfil "defensivo" en el MMPI. Su perfil se encontraba por completo dentro de límites normales, aunque había obtenido una puntuación *T* de 72 puntos en la escala K. En general se considera que la escala K es un buen índice de las actitudes defensivas durante la presentación de la prueba, en especial en valoraciones de salud mental con canalizaciones clínicas u hospitalarias. Sólo con propósitos de una breve revisión se dirá que las puntuaciones *T* en el MMPI de cerca de 50 son promedio, mientras que las elevaciones de 70 o mayores se consideran dignas de tomarse en cuenta. El psicólogo consultor señaló la calificación elevada de la candidata en la escala K, conjeturando de manera precipitada que ésta era demasiado defensiva y le advirtió al jefe de policía que no la contratara.

Lo que el psicólogo no sabía es que las calificaciones elevadas en la escala K son extremadamente comunes entre los solicitantes a un empleo dentro del sistema policiaco. Por ejemplo, Hiatt y Hargrave (1988) encontraron que cerca de 25% de una muestra de agentes policiacos produjeron perfiles del MMPI con escalas K en o por encima de una puntuación *T* de 70. De hecho, los agentes exitosos dentro de la policía tendían a tener calificaciones **¡más elevadas** en la escala K que los agentes "problemáticos"! En este caso, el usuario de la prueba no poseía los suficientes conocimientos acerca de la utilización del MMPI en la selección de personal. Su ignorancia de este elemento constituyó una violación a la ética profesional. De manera incidental, el caso se arregló fuera de los tribunales mediante una suma sustancial de dinero, lo cual demuestra que las transgresiones al uso responsable de las pruebas pueden tener graves consecuencias legales.

La destreza del psicólogo es particularmente importante cuando se utilizan los servicios de calificación e interpretación de pruebas. Los principios éticos de la *American Psychological Association* dejan poco lugar a las dudas:

> Los psicólogos conservan la responsabilidad pertinente de la aplicación, interpretación y utilización apropiadas de los instrumentos de evaluación, ya sea que ellos mismos califiquen e interpreten dichas pruebas o que utilicen servicios automatizados o de otro tipo (APA, 1992a).

Se refiere al lector al tema 15A, Evaluación psicológica auxiliada por computadoras, para una mayor discusión sobre el punto.

CONSENTIMIENTO INFORMADO

Antes de que se lleve a cabo la aplicación, el usuario debe obtener el consentimiento informado de las personas que realizarán la prueba o de sus representantes legales. En ciertos ca-

sos pueden hacerse excepciones al consentimiento informado; por ejemplo, en los programas de prueba obligatorios a nivel estatal, en las pruebas con grupos escolares y cuando el consentimiento es claramente implícito (p. ej., en pruebas para admisión a la universidad). El principio del **consentimiento informado** es tan importante que el manual de **Normas** dedica una norma específica para él:

> El consentimiento informado implica que se ha informado, en un lenguaje comprensible, a las personas que realizan la prueba o a sus representantes de las razones para la aplicación de pruebas, los tipos de éstas que se utilizarán, el uso indicado y el rango de consecuencias materiales de este uso indicado y de cuál información de las pruebas se dará a conocer y a qué personas (AERA y cols., 1985).

Incluso las personas jóvenes o con retraso mental deben recibir una explicación de las razones para la prueba. Por ejemplo, el examinador podría decir, "Voy a hacerte unas preguntas y a pedirte que trabajes con algunos problemas para ver lo que puedes hacer y encontrar en qué cosas necesitas más ayuda".

Desde un punto de vista legal, los tres elementos del consentimiento informado incluyen divulgación, competencia y participación voluntaria (Melton, Petrila, Poythress y Slobogin, 1998). El punto central de la divulgación es que el cliente reciba suficiente información (p. ej., acerca de riesgos, beneficios, difusión de informes) para tomar una decisión cuidadosa acerca de continuar con la participación en el proceso de prueba. La competencia se refiere a la capacidad mental de la persona evaluada para dar su consentimiento. En general se supone que existe competencia a menos que la persona sea un niño, un individuo muy anciano o tenga una discapacidad mental (p. ej., retraso mental). En estos casos, será necesario que un tutor proporcione consentimiento legal. Por último, la norma de participación voluntaria implica que la decisión de realizar una batería de evaluación se otorga de manera libre y no se basa en una coacción sutil (p. ej., a los internos se les promete una liberación anticipada si participan en pruebas para investigación). En la mayor parte de los casos, el examinador utiliza un formato de consentimiento informado como el que se muestra en la figura 15–3.

PRUEBAS OBSOLETAS Y ESTÁNDAR DE CUIDADO

El estándar de cuidado es un concepto indeterminado que con frecuencia surge en las revisiones profesionales o legales del ejercicio específico de profesiones relacionadas con la salud, incluyendo las pruebas psicológicas. El **estándar de cuidado** prevaleciente es aquel que es "el común, acostumbrado o razonable" (Rinas y Clyne-Jackson, 1988). Para citar un ejemplo extremo, en medicina el estándar de cuidado para una fiebre puede incluir la administración de aspirina —pero no implicaría la práctica anticuada de sangrar al paciente.

Los profesionales que utilizan pruebas psicológicas deben ser cautelosos de las pruebas obsoletas, porque su utilización podría violar el estándar de cuidado prevaleciente. Un caso a propósito es el del MMPI contra el MMPI-2. Aunque el MMPI-2 es una revisión relativamente conservadora del muy estimado MMPI, las mejorías en normas y construcción de escala son sustanciales. El MMPI-2 es el estándar de cuidado de la actualidad en la evaluación de la psicopatología con base en el sistema MMPI. Los profesionales que continúan dependiendo del MMPI original podrían estar en posibilidad de que se les demande por negligencia, en especial si la interpretación de la prueba resultó en afirmaciones interpretativas engañosas o en un diagnóstico incorrecto.

Otra cuestión relacionada con el estándar de cuidado es la utilización de resultados de una prueba que son obsoletos para el propósito actual. Después de todo, las características y rasgos individuales muestran un cambio válido a través del tiempo. Un estudiante que satisface los criterios de discapacidad para el aprendizaje en el cuarto grado de estudios, podría mostrar enormes ganancias en rendimiento académico que harían que el diagnóstico de DA ya no fuera adecuado para el quinto grado. Los

Éste es un acuerdo entre [Nombre del cliente] y [Nombre del profesional], *Ph.D.*, un psicólogo con licencia en el estado de Illinois. Se le alienta a que Ud. haga preguntas acerca de la experiencia o credenciales profesionales en cualquier momento.

1. Información general: el propósito de esta evaluación es proporcionar a su [médico, orientador, terapeuta] la información sobre su funcionamiento psicológico que pueda ser útil para el trabajo que realice con usted. La evaluación implicará una breve entrevista y pruebas psicológicas. Esto requerirá de 3 a 4 horas de su tiempo.
2. Informe de pruebas: la información relevante de la entrevista y los resultados de la prueba se resumirán en un informe por escrito que se le enviará a su [médico, orientador, terapeuta]. Los resultados de la prueba y el informe se revisarán con usted en aproximadamente una semana.
3. Confidencialidad: el informe no se proporcionará a ninguna otra fuente a menos que usted lo solicite de manera formal, por escrito. Las excepciones a esta regla incluyen las siguientes situaciones: su vida o la de otra persona está en peligro, se informa de maltrato a abuso hacia niños o ancianos o existe una orden judicial para revelar el informe.
4. Costo: se utiliza una cuota por hora de $___ para llegar a los honorarios totales. Es posible que su póliza de seguros médicos cubra parte o la totalidad de este costo. El costo total estimado de su evaluación es de $____.
5. Efectos colaterales: aunque la mayoría de los individuos disfrutan del proceso de la consulta psicológica, algunos lo consideran incómodo, en especial si los resultados de las pruebas indican problemas psicológicos. Es apropiado que usted analice sus sentimientos con el examinador. Usted está en libertad de retirar su consentimiento en cualquier momento durante el curso de las pruebas.
6. Negativa a la evaluación: usted tiene el derecho a negarse a la evaluación. No se requiere que lleve a cabo esta valoración a fin de continuar trabajando con su [médico, orientador, terapeuta]. No obstante, es más probable que su tratamiento sea efectivo si participa en esta evaluación. Si así lo solicita, analizaremos las opciones de canalización para usted.

_____ Nombre del cliente

_____ Fecha

Figura 15–3. Ejemplo abreviado del Consentimiento Informado para la Evaluación Psicológica.
Fuente: Adaptado y abreviado con autorización de Gregory, R. J. (1999). *Foundations of intellectual assessment: The WAIS-III and other tests in clinical practice.* Boston: Allyn & Bacon.

resultados en pruebas de personalidad pueden ser incluso más cambiantes. Una crisis personal a corto plazo puede causar que un perfil del MMPI-2 parezca como una cordillera montañosa. Una semana después, el perfil de esta prueba puede ser completamente normal. Es difícil proporcionar directrices en cuanto a la "vida útil" de los resultados de las pruebas psicológicas. Por ejemplo, los resultados obtenidos muchos años antes en la prueba GRE podrían todavía pronosticar de manera válida el desempeño en la universidad, mientras que los resultados en el *Beck Depression Inventory* del día de ayer podrían darle información errónea al terapeuta acerca del nivel actual de depresión. El profesional debe evaluar a nivel individual la necesidad de repetir la prueba.

REDACCIÓN RESPONSABLE DEL INFORME

Excepto por las pruebas grupales, la práctica de la evaluación psicológica culmina de manera invariable en un informe por escrito que constituye un registro semipermanente de las conclusiones de la prueba y de las recomendaciones del examinador. La redacción efectiva del informe es una habilidad importante debido al impacto potencial duradero del documento escrito. Ilustrar las cualidades de la redacción efectiva del informe va más allá del espectro de este libro, aunque sí se puede referir al lector a unas cuantas fuentes (Gregory, 1999; Tallent, 1993).

Los informes responsables utilizan de manera común una redacción sencilla y directa que se aleja de la jerga y de los tecnicismos. La meta apropiada de un informe es proporcionar perspectivas útiles acerca del cliente, ¡no impresionar a la fuente de canalización con la idea de que el examinador es una persona culta! Cuando Tallent (1993) sondeó a más de mil profesionales de la salud que hacían canalizaciones para pruebas, uno de los encuestados declaró su desprecio hacia los psicólogos que "reflejan su necesidad de brillar como un faro psicoanalítico para revelar los oscuros y profundos

secretos que han observado". En un comentario relacionado, los informes efectivos permanecen dentro de los límites del conocimiento del examinador. Por ejemplo:

> ... nunca es apropiado que un psicólogo recomiende que un cliente se someta a un procedimiento médico específico (como un rastreo de TC para un aparente tumor cerebral) o que reciba un fármaco en particular (como Prozac para la depresión). Aunque la necesidad de un procedimiento especial parezca obvia (p. ej., los síntomas reflejan de manera acentuada el rápido inicio de un padecimiento cerebral), la mejor manera de satisfacer las necesidades del cliente consiste en recomendar una consulta inmediata con el profesional médico apropiado (p. ej., neurólogo o psiquiatra) (Gregory, 1999).

En Ownby (1991) y Sattler (1988) puede encontrarse orientación adicional acerca de la redacción eficaz del informe.

COMUNICACIÓN DE LOS RESULTADOS DE PRUEBA

Los individuos que se someten a pruebas psicológicas anticipan que los resultados se les comunicarán. No obstante, es frecuente que los profesionales no incluyan una sesión individual de información de resultados como parte de la evaluación. Una de las principales razones para la renuencia es una falta de entrenamiento acerca de cómo proporcionar la información de resultados, en especial cuando éstos parecen ser negativos. Por ejemplo, ¿de qué manera le comunica un clínico a una estudiante universitaria que su CI es de 93, cuando la mayoría de los estudiantes en ese medio obtienen calificaciones de 115 o mayores?

Proporcionar información eficaz y constructiva a los clientes acerca de sus resultados de prueba es una habilidad desafiante que requiere aprendizaje. Pope (1992) enfatiza la responsabilidad del clínico para determinar que el cliente ha comprendido de manera adecuada y precisa la información que éste intentaba trasmitirle. Lo que es más, es responsabilidad del clínico el inspeccionar las reacciones adversas:

> ¿El cliente se encuentra excepcionalmente deprimido por las conclusiones? ¿El cliente infiere de las conclusiones que sugieren un trastorno del aprendizaje que —como siempre ha sospechado— es "estúpido"? Utilizar un cuidado escrupuloso para llevar a cabo esta evaluación acerca de lo que el cliente comprende y de sus reacciones ante la información de resultados no es menos importante que utilizar un cuidado adecuado en la aplicación de pruebas psicológicas estandarizadas; la aplicación de pruebas y la información de resultados son aspectos igualmente importantes, fundamentales, del proceso de evaluación (p. 271).

La información de resultados apropiada y efectiva implica un diálogo de intercambio en el que el clínico estima la manera en que el cliente ha percibido la información y busca corregir las interpretaciones potencialmente dañinas.

Es frecuente que la información de resultados tenga un efecto destructivo cuando el clínico no cuestiona las percepciones incorrectas de un cliente acerca de los resultados de prueba. Considérense en particular las pruebas de CI —un caso en el que muchas personas ponderan las calificaciones de la prueba y las consideran como un índice de valía personal. Antes de proporcionar los resultados de una prueba, se aconseja al clínico investigar lo que el cliente entiende acerca del significado de las calificaciones de CI. Después de todo, el CI es un trozo limitado del funcionamiento intelectual: no valora un impulso o carácter de algún tipo, tiene una precisión de aproximadamente ± 5 puntos, puede cambiar a través del tiempo y no evalúa muchos atributos importantes como creatividad, inteligencia social, capacidad musical o habilidad atlética. Pero un cliente puede tener una perspectiva poco realista acerca del CI y, por ello, podría llegar a conclusiones erróneas al escuchar que su calificación es de "sólo" 93. El profesional cuidadoso obtendrá del cliente sus perspectivas y las cuestionará cuando sea necesario antes de proceder. En Gass y Brown (1992) y en Pope (1992) pueden encontrarse otros conceptos acerca de la información de resultados.

Finn y Tonsager (1997) van más allá del pronunciamiento general acerca de la evitación

del daño cuando se proporciona un informe de resultados sobre pruebas, al presentar la perspectiva fascinante de que la información acerca de éstos deberían tener efectos terapéuticos directos e inmediatos para los individuos que experimentan problemas psicológicos. En otras palabras, estos autores proponen que la evaluación psicológica es una forma de intervención a corto plazo, no sólo una base para reunir información que **posteriormente** se utilizará para propósitos terapéuticos. En un estudio (Finn y Tonsager, 1992), se examinaron los efectos de una evaluación psicológica breve en los clientes de un centro de orientación universitaria. Treinta y dos estudiantes tomaron parte en una entrevista inicial, respondieron al MMPI-2 y después recibieron una sesión de una hora con la información de resultados, llevada a cabo de acuerdo con un método elaborado por Finn (1996). Un grupo de comparación de 29 estudiantes se sometió a la entrevista y recibió la misma cantidad de psicoterapia de apoyo, no directiva, en vez de la información de los resultados de la prueba. Los clientes en el grupo de evaluación con el MMPI-2 mostraron una mayor disminución en la angustia sintomática y un mayor aumento en autoestima, inmediatamente después de su sesión de información de resultados y también después de dos semanas, con respecto al grupo de comparación. El grupo de información de resultados también se mostró más esperanzado acerca de sus problemas después de la evaluación breve. Estas conclusiones ilustran la importancia de proporcionar información cuidadosa y constructiva de los resultados de prueba, más que de lanzarse a una revisión rutinaria de las conclusiones.

CONSIDERACIÓN DE LAS DIFERENCIAS INDIVIDUALES

Todas las organizaciones profesionales que tienen que ver con las pruebas psicológicas destacan el conocimiento de las diferencias individuales y el respeto hacia éstas. La *American Psychological Association* incluye el siguiente como uno de los seis principios guía:

> Principio D: respeto hacia los derechos y dignidad de las personas... Los psicólogos están conscientes de las diferencias culturales, individuales y de rol, incluyendo aquéllas debidas a la edad, género, raza, origen étnico y nacional, religión, orientación sexual, discapacidad, idioma y nivel socioeconómico. Los psicólogos intentan eliminar el efecto sobre su trabajo de los prejuicios basados en esos factores y no participan voluntariamente o toleran las prácticas discriminatorias injustas (APA, 1992a).

La importancia de este principio para las pruebas psicológicas es que se espera que los profesionales sepan cuándo una prueba o interpretación pueda no ser aplicable debido a factores como edad, género, raza, origen étnico y nacional, religión, orientación sexual, discapacidad, idioma y nivel socioeconómico. Se puede ilustrar este punto con un estudio clínico presentado por Eyde y colaboradores (1993). Un psicólogo evaluó a un hombre de 75 años de edad a petición de su esposa, quien había observado problemas de memoria. El psicólogo aplicó un examen del estado mental y una prueba de inteligencia destacada. El desempeño en el examen de estado mental era normal, pero las calificaciones en la prueba de inteligencia revelaron una fuerte discrepancia entre las subpruebas verbales y aquellas que medían capacidad espacial y velocidad de procesamiento. El psicólogo interpretó este patrón como indicación de deterioro del funcionamiento intelectual del marido. Por desgracia, esta interpretación se basó en el uso erróneo de puntuaciones estándar sin corrección para edad. Además, el psicólogo tampoco evaluó la presencia de depresión que, según se sabe, es causa de que el desempeño visoespacial tenga una disminución notable (Wolff y Gregory, 1992). De hecho, una serie de valoraciones posteriores revelaron que esta persona era un hombre perfectamente sano de 75 años de edad. El psicólogo no tomó en consideración la importancia de la edad de este señor y su situación emocional cuando interpretó la prueba de inteligencia. Éste fue un error costoso que produjo que el cliente y su esposa sufrieran de preocupaciones sustanciales innecesarias.

PRUEBAS CON MINORÍAS CULTURALES Y LINGÜÍSTICAS

ANTECEDENTES Y NOTAS HISTÓRICAS

Las personas de minorías étnicas (de origen no europeo) constituyen en la actualidad cerca de 25% de la población en EUA y se estima que abarcarán más de 50% dentro de varios decenios (Dana, 1993). No obstante, la tarea de la evaluación con pruebas se basa casi por completo en los esfuerzos de psicólogos de raza blanca que llevan una perspectiva angloestadounidense a su labor. No puede darse por sentada la idoneidad de las pruebas existentes para valorar a poblaciones diversas. La evaluación de individuos que provienen de minorías étnicas hace surgir importantes dudas, en especial cuando los resultados de una prueba se traducen en decisiones de asignación o en otros resultados delicados, como ocurre por lo común dentro de instituciones educativas.

Como se señaló en el capítulo 1, Historia de las pruebas psicológicas, los pioneros en el movimiento de las pruebas ignoraron en gran medida el impacto de los antecedentes culturales sobre los resultados de éstas. Por ejemplo, en el decenio de 1920, Henry Goddard concluyó que la inteligencia del inmigrante promedio era alarmantemente baja, "quizás al grado de morón". Sin embargo, le restó importancia a la probabilidad de que las diferencias lingüísticas y culturales pudieran explicar las bajas calificaciones en una prueba de los inmigrantes.

Quizá como un rechazo en contra de estos primeros métodos, a inicios de la década de 1930 los psicólogos presentaron una mayor sensibilidad a las variables culturales en la práctica de las pruebas. Un resplandeciente ejemplo en este sentido fue Stanley Porteus, quien emprendió una investigación de gran alcance acerca del temperamento e inteligencia de los pueblos aborígenes australianos. Porteus (1931) utilizó muchos instrumentos tradicionales (diseños con cubos, laberintos, retención de dígitos) pero para crédito de este investigador, también diseñó una medida ecológicamente válida de la inteligencia para este grupo; a saber, reconocimiento de huellas de pisada. Mientras que los aborígenes evaluados tuvieron un desempeño deficiente en las pruebas eurocéntricas, su capacidad para reconocer huellas en fotografía estaba a la par con otros grupos raciales estudiados. Aun así, Porteus se mostró muy consciente en cuanto a que su procedimiento **todavía** podría representar un impedimento para los aborígenes:

> La fotografía de una huella de pisada no es lo mismo que la pisada en sí y es muy probable que varias señales que utiliza el rastreador aborigen estén ausentes en una fotografía. Las profundidades diversas de partes de la impresión del pie no son visibles en la fotografía y puede ser que las peculiaridades individuales, diferentes a la forma y tamaño generales de la huella, no resalten de manera clara. De aquí que se deba esperar que los individuos aborígenes se encuentren en cierta desventaja al equiparar estas fotografías de huellas con el reconocimiento de las huellas en sí (pp. 399-400).

De manera similar, DuBois (1939) encontró que los niños de los indígenas pueblo presentaban una capacidad superior en la prueba del dibujo de un caballo, que el autor diseñó de manera especial para evaluar la capacidad mental de estos niños, mientras que su desempeño era menos adecuado en la prueba tradicional de Goodenough (1926) del Dibujo de la Figura Humana. Desde estos primeros estudios, los psicólogos han mantenido un especial interés en el impacto del lenguaje y la cultura en el significado de los resultados de una prueba.

IMPACTO DE LOS ANTECEDENTES CULTURALES EN LOS RESULTADOS DE UNA PRUEBA

Los profesionales necesitan apreciar que los antecedentes culturales de las personas evaluadas tendrán un impacto en todo el proceso de evaluación. Por esta razón, Sattler (1988) aconseja que los psicólogos dedicados a la evalua-

ción enfoquen su tarea desde una perspectiva plural:

> Los grupos culturales pueden variar con respecto a valores culturales (lo cual se deriva en parte del choque, discontinuidad o conflicto culturales); lenguaje y matices en el estilo del lenguaje; perspectivas acerca de la vida y de la muerte; roles de los miembros de la familia; estrategias de solución de problemas; actitudes hacia la educación, salud mental y enfermedad mental; y etapa de aculturación (el grupo puede seguir valores tradicionales, aceptar los valores del grupo dominante o estar en algún punto entre ambos). El examinador debe adoptar un marco de referencia que le permita comprender cómo ciertas conductas tienen sentido dentro de cada cultura (p. 565).

A pesar de reconocer el impacto de las diferencias culturales sobre las pruebas, también es importante evitar las generalizaciones estereotípicas excesivas. Por ejemplo, con frecuencia se señala que los nativos estadounidenses presentan un concepto distintivo acerca del tiempo, enfatizando el **presente** en contraposición a la orientación hacia el **futuro** que es tan poderosamente formativa en la sociedad estadounidense de raza blanca, de clase media (Panigua, 1994). Una posible implicación de esta diferencia cultural es que los límites de tiempo podrían no representar lo mismo para un niño indígena estadounidense que para uno de la cultura predominante. Quizás el niño de minoría desobedezca las instrucciones de una subprueba y trabaje a un paso cuidadoso y medido, en vez de buscar las soluciones rápidas. Por supuesto, es posible que este niño obtenga entonces una calificación engañosamente baja en esa medida.

No obstante, al afirmar las diferencias culturales, los examinadores necesitan evitar el peligro de los estereotipos. La cultura no es monolítica. Cada persona es única. Algunos indígenas estadounidenses exhibirán una orientación distintiva acerca del tiempo, pero quizá la mayoría no lo hagan. El reto para el profesional consiste en observar los detalles clínicos del desempeño e identificar los matices culturales de la conducta que ayudan a determinar los resultados de una prueba.

Un ingenioso estudio de Moore (1986) ilustra de manera poderosa la importancia de los antecedentes culturales para comprender el desempeño en una prueba de los individuos de minorías étnicas. Esta autora no sólo comparó las calificaciones en pruebas de inteligencia, sino también **la manera cualitativa** de responder a las demandas de la prueba de dos grupos de niños adoptivos afroestadounidenses. Un grupo lo constituían 23 niños adoptados de modo transracial por familias de raza blanca, de clase media. El otro lo formaban 23 niños adoptados de manera intrarracial por familias afroestadounidenses. A todos los niños se les había adoptado antes de los dos años de edad y los antecedentes de las familias adoptivas eran similares en términos de educación y clase social. Así, las diferencias en calificaciones y conductas de prueba podían atribuirse principalmente a aquéllas en antecedentes culturales que surgían del hecho de que un grupo había sido adoptado por familias afroestadounidenses, mientras que al otro lo habían adoptado familias de raza blanca. Las pruebas y observaciones las llevaron a cabo dos examinadoras afroestadounidenses, "ciegas" acerca del propósito del estudio. Los niños adoptados de manera transrracial y examinados de los 7 a los 10 años de edad obtuvieron un CI promedio de 117 en el WISC, en comparación con un CI promedio de 104 para los niños adoptados de manera tradicional. Estos resultados de CI no fueron notables, en la medida en que Scarr y Weinberg presentaron conclusiones similares años antes.

El sorprendente e informativo resultado fue que los dos grupos de niños mostraron conductas **cualitativas** muy diferentes durante la prueba. Como grupo, los niños con calificaciones menores de CI (aquéllos adoptados por familias afroestadounidenses) estaban en menor probabilidad de elaborar de manera espontánea sus respuestas de trabajo y en mayor probabilidad de simplemente negarse a responder cuando se les presentaba una demanda de la prueba. Moore (1986) ofreció las siguientes interpretaciones:

> La tendencia de los niños a elaborar de manera espontánea sus respuestas de trabajo puede ser

> un índice muy importante de su nivel de participación en el desempeño de la tarea, sus estrategias de solución de problemas, nivel de motivación para generar una respuesta correcta y nivel de adaptación a la situación estandarizada de prueba[...] Aunque la respuesta terminal no trabajada se considera como incorrecta, de hecho no proporciona ninguna documentación empírica de que lo que el niño sabe o no, o de lo que el niño puede o no puede hacer. La única información disponible es que el niño no responde a la demanda (p. 322).

La lección esencial de este estudio es que las diferencias con bases culturales en cuanto a estilo de respuesta pueden tener por resultado la ocultación de la capacidad subyacente de algunas personas evaluadas. Siempre es aconsejable hacer interpretaciones cuidadosas de los resultados de una prueba, pero esto tiene particular importancia en el caso de individuos que provienen de antecedentes cultural o lingüísticamente diversos.

La influencia de los factores culturales no se limita al desempeño de los niños en una prueba, sino que también abarca a los adultos. Terrell, Terrell y Taylor (1981) investigaron los efectos de la confianza/desconfianza racial sobre las calificaciones en pruebas de inteligencia de estudiantes universitarios afroestadounidenses. Los investigadores identificaron a los estudiantes afroestadounidenses con niveles altos y bajos de desconfianza hacia los de raza blanca. Con un diseño de 2×2, a la mitad de cada grupo se le sometió a una prueba individual de inteligencia aplicada por un examinador de raza blanca, mientras que a la otra mitad un examinador afroestadounidense les aplicó la prueba. Como se pronosticó, el análisis de varianza no reveló diferencias en los efectos principales de la raza del examinador (blanca contra afroestadounidense) o del nivel de desconfianza (alto contra bajo; figura 15–4). Pero se reveló una interacción sustancial; es decir, el grupo de alta desconfianza y con un examinador afroestadounidense obtuvo calificaciones mucho mejores que el grupo de alta desconfianza con un examinador de raza blanca (CI promedio de 96 contra 86, respectivamente). Dicho de manera sencilla, la desconfianza cultural entre los afroestadounidenses se asoció con calificaciones significativamente menores de CI, pero **sólo** cuando el examinador era de raza blanca.

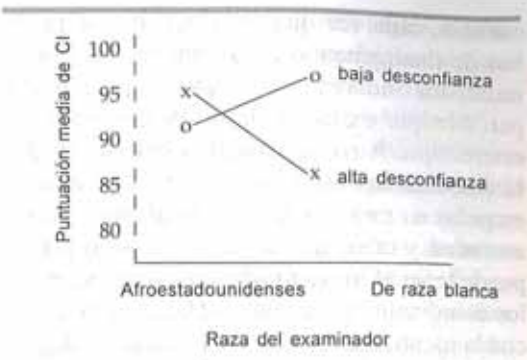


Figura 15–4. Puntuaciones medias de CI de estudiantes afroestadounidenses en función de la raza del examinador y de la desconfianza cultural.
Fuente: Basado en datos de Terrell, F., Terrell, S. y Taylor, J. (1981). Effects of race of examiner and cultural mistrust on the WAIS performance of Black students. *Journal of Consulting and Clinical Psychology*, 49, 750-751.

Para ilustrar aún más las influencias culturales, Steele (1997) ha propuesto una teoría acerca de que los estereotipos de la sociedad sobre los grupos influyen sobre el desempeño intelectual inmediato y también el desarrollo a largo plazo de la identidad de los miembros individuales del grupo. Steele ha aplicado esta teoría tanto a mujeres —donde los estereotipos afectan su aprovechamiento en matemáticas y ciencias naturales— como a afroestadounidenses —donde los estereotipos aparentemente disminuyen su desempeño en pruebas estandarizadas. Aquí se analiza su investigación sobre la amenaza del estereotipo con estudiantes universitarios afroestadunidenses (Steele y Aronson, 1995).

La idea de la **amenaza del estereotipo** es, en esencia, una versión compleja de una profecía autocumplida. Los investigadores la definen como la amenaza de confirmar, como característica propia, un estereotipo negativo acerca del propio grupo. Por ejemplo, con base en datos publicados y la cobertura de medios de información acerca de la raza y las calificaciones de CI, se ha estereotipado a los afroestadounidenses como poseedores de menor capacidad intelectual que los demás. En conse-

cuencia, cada vez que se enfrentan con pruebas de inteligencia o aprovechamiento académico, los individuos de este grupo pueden percibir que existe el riesgo de confirmar el estereotipo. A corto plazo, la hipótesis es que la amenaza del estereotipo disminuirá el desempeño en una prueba debido al aumento en ansiedad y otros mecanismos. A largo plazo, puede tener el impacto adicional de presionar a los estudiantes afroestadounidenses a "romper con la identificación como manera de protegerse" al obtener logros en la escuela y áreas intelectuales relacionadas.

Steele y Aronson (1995) llevaron a cabo una serie de cuatro estudios para valorar la hipótesis de la amenaza del estereotipo. Todas las investigaciones confirmaron la hipótesis. En este caso se enfoca la atención en el primer estudio, en el que se aplicó una prueba de 30 minutos a estudiantes afroestadounidenses y de raza blanca, compuesta por reactivos difíciles de la sección verbal del *Graduate Record Examination* (Examen de registro del posgrado). Se asignó de manera aleatoria a los estudiantes de ambos grupos raciales a una de tres condiciones de prueba: amenaza del estereotipo, en la que la prueba se describió como un diagnóstico de la capacidad verbal individual; control, en la que ésta fue descrita sólo como herramienta de investigación; y control-desafío, en la que se le describió sólo como herramienta de investigación, pero a los participantes se les exhortó a "asumir el desafío de manera seria". Las calificaciones en la prueba verbal se ajustaron (análisis de covarianza) con base en las puntuaciones previas de aprovechamiento, a fin de eliminar los efectos de diferencias preexistentes entre grupos.

Las diferencias raciales fueron pequeñas y no significativas en las condiciones control y control-desafío, mientras que los individuos afroestadounidenses obtuvieron calificaciones mucho más bajas que los de raza blanca en la condición de amenaza del estereotipo (figura 15–5). En otros estudios, Steele y Aronson (1995) investigaron el mecanismo de mediación a través del cual la amenaza del estereotipo provocaba que los afroestadounidenses tuvieran calificaciones más bajas en pruebas estandarizadas. Los detalles van más allá del espectro de este libro, pero la conclusión general no:

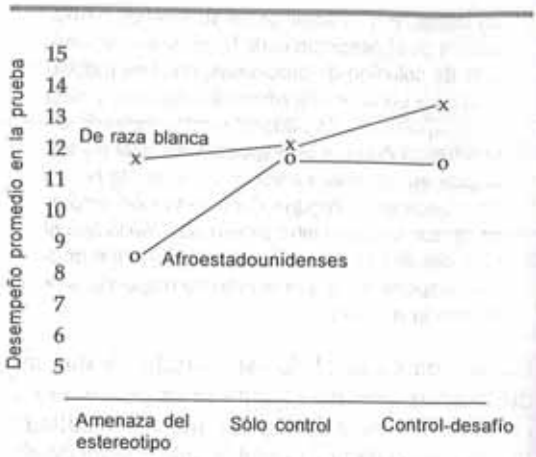


Figura 15–5. Promedio de reactivos verbales correctos para personas de raza blanca y afroestadounidenses bajo tres condiciones.
Fuente: Basado en datos de Steele, C. M. y Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*, 797-811.

> Nuestra mejor evaluación es que la amenaza del estereotipo causa una insuficiencia del procesamiento muy parecida a la provocada por otras presiones evaluativas. Los participantes con amenaza del estereotipo ocuparon más tiempo en responder a menos reactivos de manera más imprecisa —probablemente como resultado de alternar su atención entre tratar de responder a los reactivos y de evaluar el significado para sí mismos de su frustración (Steele y Aronson, 1995, 809).

En resumen, los autores proponen una perspectiva sociopsicológica acerca del significado de las menores calificaciones de prueba de los afroestadounidenses y quizá también de otros grupos amenazados por el estereotipo. Su punto de vista enfatiza que los resultados de una prueba no residen dentro de los individuos. Las calificaciones de una prueba ocurren dentro de un complejo campo sociopsicológico en el que influyen potencialmente la historia nacional, los conflictos de raza y muchos otros factores sutiles.

EVALUACIÓN DE MINORÍAS CULTURALES Y LINGÜÍSTICAS

En los últimos 50 años, es cada vez más frecuente que el campo de la psicología haya reconocido que puede ser necesario utilizar prácticas especializadas para lograr un examen equitativo con las minorías culturales y lingüísticas. La sensibilidad a las cuestiones únicas que se enfrentan en la evaluación de minorías surgió por vez primera en el decenio de 1960 cuando la *Society for the Psychological Study of Social Issues* publicó las pautas para examinar a niños de minorías. De manera más reciente, la profesión de la psicología ha reiterado su preocupación en cuanto a la validez de la evaluación de minorías culturales y lingüísticas al publicar pautas para quienes proporcionan servicios psicológicos a poblaciones minoritarias (APA, 1993). Estas pautas amplias proporcionan metas que representan una aspiración (p. ej., se espera que los examinadores reconozcan la diversidad cultural), pero que representan muy poco en el sentido de consejos específicos para el profesional de la evaluación. Como resultado ha aparecido una oleada de influyentes libros y artículos, cada uno de los cuales ofrece comentarios cuidadosos acerca de la evaluación multicultural (p. ej., Lam, 1993; Suzuki, Meller y Ponterotto, 1996; Rogers, 1998). Como se expresa a continuación, en esta literatura se pueden identificar ciertos temas aleccionadores.

Una consideración preponderante es que las barreras lingüísticas pueden inhibir el desempeño en una prueba de los individuos de minorías. Valdes y Figueroa (1994) expresan el problema de la siguiente manera:

> Cuando un individuo bilingüe se enfrenta con una prueba monolingüe, desarrollada por individuos monolingües y cuyas estandarización y normas se establecieron con una población monolingüe, tanto a la persona que contesta la prueba como a la prueba en sí se les pide hacer algo que no pueden. La persona bilingüe que presenta la prueba no puede tener el mismo desempeño que un individuo monolingüe. La prueba monolingüe no puede "medir" en el otro idioma (p. 172).

Lo que esto significa desde un punto de vista práctico es que, incluso aunque un niño bilingüe de minoría tenga fluidez en inglés, aun así una prueba en este idioma podría subestimar su verdadero nivel de capacidad. El examinador sensible reconocerá esta posibilidad al interpretar los resultados de la prueba.

Podría parecer que existe la posibilidad de utilizar a un intérprete de la lengua materna para facilitar la evaluación de las personas cuyo idioma principal no sea el inglés. En general, los especialistas en pruebas no aconsejan esta práctica porque los intérpretes pueden sustituir palabras, hablar en un dialecto diferente o dar sugerencias sutiles que influyen las respuestas de los individuos evaluados (Rogers, 1998). Sería preferible un psicólogo bilingüe bien entrenado, pero incluso esta práctica les parece problemática a algunos (Figueroa, 1990).

La opción preferida consiste en utilizar pruebas traducidas a la lengua materna de la persona evaluada y cuyas normas se hayan establecido en subpoblaciones relacionadas. Por desgracia, existen relativamente pocos instrumentos disponibles para este propósito. Las versiones en español de pruebas destacadas son la excepción a esta situación. Algunos ejemplos incluyen la Escala de Inteligencia Wechsler para Niños–Revisada de Puerto Rico (EIWN-R PR), una versión en español del WISC-R, y la prueba TEMAS, una versión en español del Test de Apercepción Temática que se compone de láminas estímulo nuevas y relacionadas culturalmente. La EIWN-R PR es una adaptación al español del WISC-R realizada en 1993. Las normas de esta prueba se obtuvieron con 2 200 niños puertorriqueños entre los 6 y 16 años, 11 meses de edad. La EIWN-R PR conserva la estructura y contenido originales del WISC-R, además de proporcionar instrucciones en español para la aplicación y calificación de las 12 subpruebas. El TEMAS se analizó de manera breve en el tema 13B, Técnicas proyectivas.

Además de las posibles barreras lingüísticas, los individuos de minorías pueden exhibir una falta de elaboración para presentar pruebas, lo cual agrava todavía más su situación desventajosa. Desde una perspectiva hispana,

Padilla y Medina (1996) hacen las siguientes observaciones:

> Es bastante probable que los estudiantes de minorías se encuentren menos familiarizados con las pruebas estandarizadas de rendimiento y, por tanto, tengan menos habilidades asociadas con responder pruebas que los estudiantes de mayoría, quienes se han visto expuestos a las pruebas estandarizadas durante un tiempo largo. Otra consideración es que los padres más educados, que tienen conocimientos acerca de cómo responder a las pruebas, participen más en un entrenamiento de sus hijos para instruirlos sobre las estrategias que se conocen como útiles en las pruebas de opción múltiple y sobre la importancia de equilibrar la velocidad y la precisión en pruebas de tipo objetivo. Por lo general, este tipo de entrenamiento y práctica no se encuentra en los hogares de niños de nivel socioeconómico más bajo, debido a que los padres pueden ignorar las prácticas adecuadas que se asocian con presentar pruebas (p. 14).

No existe otro remedio inmediato para este tipo de problema que reconocer que las calificaciones en una prueba de los individuos de minorías **pueden** reflejar una falta de elaboración hacia las pruebas.

La probabilidad de que las barreras lingüísticas y la falta de elaboración hacia las pruebas influya en los resultados de las minorías es un fuerte argumento en favor de utilizar un enfoque **multidisciplinario** para la evaluación. Cuando la evaluación representa las perspectivas múltiples de varias disciplinas (p. ej., especialistas en psicología, habla y lectura) es menos probable que los resultados erróneos de la evaluación que provengan de cualquier disciplina por sí sola deriven en un perjuicio. Esto es particularmente cierto en la evaluación escolar de niños que provienen de antecedentes culturales y lingüísticos diferentes. Rogers (1998) recomienda que la evaluación de niños de minorías proceda de las siguientes maneras:

1. Evaluaciones multidisciplinarias que impliquen la obtención de información a partir de una variedad de fuentes y métodos;
2. Evaluaciones llevadas a cabo en la lengua materna del niño, así como en inglés;
3. Evaluaciones que protejan a los niños de las prácticas de selección y administración que sean discriminatorias en sentido racial y cultural;
4. Procedimientos especificados de manera clara para evaluar a niños de antecedentes lingüísticos diferentes; y
5. Consentimiento informado de los padres y notificación a los mismos acerca de sus derechos a los procedimientos legales establecidos.

Estos asuntos nos recuerdan una vez más que la práctica de la evaluación ocurre dentro de un contexto social cargado de valores. El examinador capaz presenta una sensibilidad hacia la diversidad cultural y lingüística en la aplicación e interpretación de pruebas psicológicas.

REITERACIÓN: USO RESPONSABLE DE LAS PRUEBAS

Se regresa ahora a los verdaderos dilemas de la aplicación de pruebas que se mencionaron al inicio del tema. El lector recordará que el primer dilema tenía que ver con el hecho de si un psicólogo responsable puede negarse a dar información sobre resultados de los candidatos a agentes de la policía que se canalizan para un examen de selección previo a su contratación. De manera sorprendente, la respuesta para esta pregunta es Sí. Bajo circunstancias normales, un profesional debe explicarle los resultados de la evaluación al cliente. Pero existen excepciones, como lo explica el Principio 2.09 del Código Ético de la APA:

> A menos que, de antemano, se explique con claridad a la persona evaluada la naturaleza de la relación y se limite la entrega de resultados (como en algunas valoraciones de consultoría organizacional, selección previa a la contratación o de seguridad y evaluaciones forenses), los psicólogos garantizan que se proporcionará una explicación de resultados en la que se utilice un lenguaje que razonablemente pueda comprender la persona evaluada u otra que tenga autori-

zación legal para representar al cliente (APA, 1992a).

El segundo dilema tenía que ver con un orientador que continuaba utilizando el MMPI aunque el MMPI-2 estaba disponible desde hacía varios años. ¿La negativa del orientador a utilizar el MMPI-2 es un muestra de incumplimiento de las normas profesionales? La respuesta a esta pregunta es probablemente Sí. El MMPI-2 tiene validación adecuada y constituye una mejoría significativa con respecto al MMPI. Como se mencionó antes, el MMPI-2 es ahora el estándar de cuidado en la evaluación de la psicopatología con base en el sistema MMPI. El orientador que siguiera utilizando el MMPI original estaría en posibilidad de que se le demandara por negligencia, en especial si sus interpretaciones para la prueba resultaran en afirmaciones interpretativas engañosas o en un diagnóstico falso.

El tercer problema implicó a alguien como traductor en la aplicación del WISC-III a un niño de nueve años de edad cuyo primer idioma era el español. En general, éste es un error, ya que sacrifica el control estricto de los materiales de la prueba. El examinador no era bilingüe y, por tanto, no tenía manera de saber si el traductor era fiel al texto original o posiblemente estaba proporcionando indicios adicionales. En un mundo ideal, el procedimiento apropiado consistiría en incorporar a un examinador hispanohablante que utilizara una prueba traducida de manera formal y también estandarizada con personas de origen hispano. Por ejemplo, la Escala de Inteligencia Wechsler para Niños–Revisada de Puerto Rico (EIWN-R PR), que se analizó anteriormente, sería una buena opción.

El último dilema tenía que ver con la cliente, que le informó a un psicólogo que su hermano recientemente muerto había sido, con toda probabilidad, un pedófilo. ¿El psicólogo está obligado a informar de este caso a las autoridades competentes? Sí, pero ello podría depender de la jurisdicción del psicólogo y del fraseo de los estatutos pertinentes. De hecho, el psicólogo sí informó del caso a las autoridades, con inesperadas consecuencias. La policía obtuvo una orden de cateo, fue al hogar de la madre de la cliente (donde el hermano había vivido) y registró la habitación de éste. La madre se traumatizó a causa de la inesperada visita de la policía y culpó del infortunio a su hija. A raíz de ello se dio una separación y la cliente demandó entonces al psicólogo por ¡violación de la confidencialidad!

RESUMEN

1. Como ocurre con todas las actividades profesionales de los psicólogos, la aplicación de pruebas se guía por normas éticas y profesionales. El empleo responsable de pruebas se define por pautas escritas y publicadas por asociaciones profesionales como la *American Psychological Association* y otros grupos.
2. Los editores de pruebas también siguen pautas profesionales que incluyen la expectativa de que venderán pruebas de alta calidad, promoverán sus productos de manera responsable y restringirán su distribución sólo a las personas con las credenciales apropiadas.
3. Aunque existen excepciones, en general la aplicación de pruebas se guía por una pregunta suprema: ¿cuáles son los intereses del cliente? La implicación funcional de esta pauta es que la evaluación debería satisfacer un propósito constructivo para la persona examinada.
4. Los psicólogos tienen la primordial obligación de garantizar la confidencialidad de la información, incluyendo resultados de una prueba, que obtengan de sus clientes en el

curso de las consultas. Algunas excepciones incluyen circunstancias poco comunes en las que retener la información representaría un claro peligro para el cliente o para otras personas.

5. Los psicólogos tienen el deber de advertir que se deriva de la decisión de 1976 para el caso Tarasoff. Los clínicos deben comunicar cualquier amenaza grave, tanto a una víctima potencial como a las autoridades competentes.
6. La responsabilidad última de la aplicación apropiada de las pruebas siempre reside en el usuario de las mismas. Desde un punto de vista práctico, esto significa que el usuario de pruebas debe tener el entrenamiento adecuado en evaluación y teoría de la medición.
7. La norma profesional acerca del consentimiento informado dispone que se debe informar a las personas examinadas acerca de las razones de la prueba, los tipos de pruebas a utilizar, las posibles consecuencias de la aplicación de las mismas y de cuál información acerca de éstas se dará a conocer y a quién.
8. El estándar de cuidado prevaleciente es el común, acostumbrado y razonable. Satisfacer el estándar de cuidado significa que el psicólogo debe abstenerse de utilizar pruebas anticuadas, en especial cuando está disponible una nueva edición.
9. Otras pautas para el empleo responsable de pruebas incluyen la redacción cuidadosa y eficaz del informe, así como proporcionar a los individuos evaluados un reporte prudente y sensible de resultados que aclare sus errores conceptuales.
10. Otra expectativa es que la evaluación se guiará por el conocimiento y respeto hacia las diferencias individuales. Por ejemplo, los profesionales deben conocer los efectos de la edad, género, raza, origen étnico y otras variables asociadas con antecedentes sobre los resultados de una prueba.
11. Los factores culturales que pueden influir en los resultados de una prueba incluyen la manera **cualitativa** de enfocar una prueba, la confianza/desconfianza racial y la amenaza del estereotipo, que es la amenaza de confirmar, como característica propia, un estereotipo negativo acerca del propio grupo.
12. Las barreras lingüísticas también pueden inhibir el desempeño en una prueba de los individuos de minorías. Las personas bilingües, y los individuos cuya lengua materna no es el inglés, pueden enfrentar problemas sutiles en las pruebas desarrolladas para utilizarse en la cultura predominante.
13. Una falta de elaboración acerca de la naturaleza de las pruebas es otro factor que enfrentan algunos individuos de minoría. Las barreras del idioma y la falta de elaboración hacia las pruebas son fuertes argumentos a favor de utilizar un enfoque **multidisciplinario** para la evaluación (p. ej., especialistas en psicología, habla y lectura).

TÉRMINOS Y CONCEPTOS CLAVE

Obligación de advertir p. 691
Consentimiento informado p. 693
Estándar de cuidado p. 693
Amenaza del estereotipo p. 699

Principales hitos en la historia de las pruebas psicológicas

APÉNDICE A

2200 a.C.	Los chinos comienzan los primeros exámenes para el servicio público.
1838	Jean Esquirol distingue entre enfermedad mental y retraso mental.
1862	Wilhelm Wundt utiliza un péndulo calibrado para medir la "velocidad del pensamiento".
1866	O. Edouard Seguin escribe el primer libro de texto importante acerca de la evaluación y tratamiento del retraso mental.
1869	Wundt funda el primer laboratorio de psicología en Leipzig, Alemania.
1884	Francis Galton aplica la primera batería de pruebas a miles de ciudadanos en la Muestra Internacional de Salud.
1890	James McKeen Cattell utiliza el término **prueba mental** cuando anuncia sus planes para su batería galtoniana de pruebas.
1896	Emil Kraepelin proporciona la primera clasificación amplia de los trastornos mentales.
1901	Clark Wissler descubre que los "instrumentos de bronce" de las prue-

bas cattellianas no tienen correlación con las calificaciones escolares.

1904 Charles Spearman propone que la inteligencia consiste de un sólo factor general, *g*, y numerosos factores específicos, s_1, s_2, s_3, etcétera.

1904 Karl Pearson formula la teoría de la correlación.

1905 Alfred Binet y Theodore Simon inventan la primera prueba moderna de la inteligencia.

1908 Henry H. Goddard traduce las escalas Binet-Simon del francés al inglés.

1912 Stern introduce el CI o cociente de inteligencia: la edad mental dividida entre la edad cronológica.

1916 Lewis Terman revisa las escalas Binet-Simon, publica la Stanford-Binet; las revisiones de esta prueba aparecen en 1937, 1960 y 1986.

1917 Robert Yerkes encabezó el desarrollo de las pruebas Army Alfa y Beta utilizadas en los exámenes para los reclutas de la primera Guerra Mundial

1917 Robert Woodworth desarrolló la Hoja de datos personales, la primera prueba de personalidad.

1920 Se publica la prueba de interpretación de Formas de Rorschach.

1921 Cattell, Thorndike y Woodworth fundan la *Psychological Corporation* —primera editorial importante de pruebas.

1926 Florence Goodenough publica la Prueba de la figura humana.

1926 La Junta de examen de ingreso a la universidad publica la primera *Scholastic Aptitude Test* (Prueba de Aptitudes Escolares).

1927 Se publica la primera edición del *Strong Vocational Interest Blank* (Formulario Strong de Intereses Vocacionales).

1935 Morgan y Murray en la Universidad de Harvard publican el Test de apercepción temática.

1936 Lindquist y otros autores publican el precursor de las *Iowa Tests of Basic Skills* (Pruebas Iowa de Habilidades Básicas).

1936 Edward Doll publica la *Vineland Social Maturity Scale* (Escala Vineland de Madurez Social) para la evaluación de conducta adaptativa en individuos con retraso mental.

1938 L. L. Thurstone propone que la inteligencia consiste de aproximadamente siete factores de grupo conocidos como habilidades mentales primarias.

1938 Raven publica las Matrices progresivas de Raven, una prueba no verbal cuyo propósito es medir el factor *g* de Spearman.

1938 Lauretta Bender publica el Test Guestáltico Visomotor, una prueba de copiado de diseños acerca de la integración visomotora.

1938 Oscar Buros publica el primer *Mental Measurements Yearbook*.

1938 Arnold Gesell da a conocer su escala de desarrollo infantil.

1939 Se publica la Escala Wechsler-Bellevue de Inteligencia; las revisiones se publican en 1955 (WAIS), 1981 (WAIS-R) y 1997 (WAIS-III).

1939 Publicación de las tablas Taylor-Russell para determinar la proporción esperada de solicitantes exitosos a través de una prueba.

1939 Se publica el Registro de preferencias de Kuder, un inventario de intereses de opción forzosa.

1942 Se publica el Inventario Multifásico de Personalidad de Minnesota (MMPI).

1948 La *Office of Strategic Services* (OSS) utiliza técnicas situacionales para la selección de oficiales del ejército.

1949 Se publica la Escala Wechsler de inteligencia para niños; aparecen revisiones en 1974 (WISC-R) y 1991 (WISC-III).

1950 Aparece el *Rotter Incomplete Sentences Blank* (Formulario de Oraciones incompletas de Rotter).

1951 Lee Cronbach introduce el coeficiente alfa como índice de la confiabilidad (consistencia interna) para pruebas y escalas.

1952 La *American Psychiatric Association* publica el **Manual Diagnóstico y Estadístico (DSM I)**.

1953 Stephenson desarrolla la técnica Q para estudiar el autoconcepto y otras variables.

1954 Paul Meehl publica *Clinical vs. Statistical Prediction*

1956 Comienza a surgir la *Halstead-Reitan Test Battery* (Batería Halstead-Reitan) como la primera batería de pruebas en neuropsicología.

1957 C. E. Osgood describe el diferencial semántico.

1958 Lawrence Kohlberg publica la primera versión de su Escala de juicio moral; la investigación con ella se expande hasta mediados del decenio de 1980.

1959 Campbell y Fiske publican un enfoque de validación de pruebas que se conoce como matriz multirrasgo-multimétodo.

1963 Raymond Cattell propone la teoría de las inteligencias fluida y cristalizada.

1967 En el caso de **Hobson contra Hansen**, el tribunal resuelve en contra del uso de las pruebas grupales de capacidad para "detectar" estudiantes, con base en que dichas pruebas discriminan en contra de niños de minorías.

1968 La *American Psychiatric Association* publica el **DSM-II**.

1969 Nancy Bayley publica las *Bayley Scales of Infant Development* (BSID), (Escalas Bayley de Desarrollo Infantil). En 1993 se publica la versión revisada (BSID-2).

1969 Arthur Jensen propone la hipótesis genética de las diferencias de CI entre personas afroestadounidenses y de raza blanca en la *Harvard Educational Review*.

1971 En el caso de **Griggs contra Duke Power**, la Suprema Corte establece que los resultados de las pruebas para obtener un empleo deben tener una vinculación demostrable con el desempeño laboral.

1971 George Vaillant populariza una jerarquía de 18 mecanismos adaptativos del yo y describe una metodología para su evaluación.

1971 Una decisión de la corte requiere que las pruebas utilizadas para selección de personal deben relacionarse con el puesto (**Griggs contra Duke Power**).

1972 Se publica la regla del Código Penal Modelo para la insanidad legal y se le adopta de manera amplia en EUA.

1974 Rudolf Moos comienza la publicación de las Escalas de clima social para evaluar diferentes ambientes.

1974 Friedman y Rosenman popularizan el patrón conductual Tipo A, con propensión a la cardiopatía; su evaluación se basa en entrevistas.

1975 El Congreso de EUA aprueba la Ley pública 94-142, el Decreto de educación para todos los niños con discapacidad.

1978 Jane Mercer publica el SOMPA *(System of Multicultural Pluralistic Assessment*; Sistema de evaluación

multicultural pluralista), una batería de pruebas diseñada para reducir la discriminación cultural.

1978 En las *Uniform Guidelines on Employee Selection*, el impacto adverso se define a través de la regla de las cuatro quintas partes; también se publican los estudios sobre pautas para la selección de personal.

1979 En el caso de **Larry P. contra Riles**, el tribunal establece que las pruebas estandarizadas de CI muestran sesgo cultural en contra de niños de raza negra con funcionamiento inferior.

1980 En el caso de *Parents in Action on Special Education* **contra Hannon**, el tribunal resuelve que las pruebas estandarizadas de CI no están sesgadas a nivel racial o cultural.

1985 La *American Psychological Association* y otros grupos publican de manera conjunta los influyentes *Standards for Educational and Psychological Testing*.

1985 Sparrow y otros autores publican las *Vineland Adaptive Behavior Scales* (Escalas Vineland de Conducta Adaptativa), una revisión de la *Vineland Social Maturity Scale* de 1936, una prueba líder en el área.

1987 La *American Psychiatric Association* publica el **DSM-III-R.**

1989 Se señala el "Efecto del Lago Wobegon". Casi todos los estados de la EUA afirman que sus niveles de rendimiento están por encima del promedio.

1989 Se publica el Inventario multifásico de la personalidad de Minnesota-2

1992 La *American Psychological Association* publica una revisión de los *Ethical Principles of Psychologist and Code of Conduct* (*American Psychologists*, diciembre de 1992).

1994 La *American Psychiatric Association* publica el **DSM-IV.**

1994 Herrnstein y Murray reviven el debate de la raza y la hereditabilidad del CI en *The Bell Curve.* ■

APÉNDICE B

Direcciones de editoriales de prueba

American Association on Mental Retardation
444 North Capitol Street, N. W., Suite 846
Washington, DC 20001-1512

American College Testing Program
P.O. Box 168
Iowa City, IA 52243-0168

American Guidance Service
P.O. Box 99
Circle Pines, MN 55014-1796

C.P.S., Inc.
P.O. Box 83
Larchmont, NY 10538

Consulting Psychologists Press
P.O. Box 10096
Palo Alto, CA 94303

CTB/Macmillan/McGraw-Hill
20 Ryan Ranch Road
Monterey, CA 93940

Denver Developmental Materials
P.O. Box 20037
Denver, CO 80220

DLM Teaching Resources
One DLM Park
Allen, TX 75002

Educational and Industrial Testing Service (EdITS)
P.O. Box 7234
San Diego, CA 92167

Educational Testing Service
ETS Test Collection (30-B)
Rosedale Road
Princeton, NJ 08541-001

GED Testing Service
One Dupont Circle, N. W.
Washington, DC 20036

Harvard University Press
79 Garden Street
Cambridge, MA 02138

Hawthorne Educational Services
800 Gray Oak Drive
Columbia, MO 65201

Marshal S. Hiskey
5640 Baldwin
Lincoln, NE 68507

Hogrefe & Huber Publishers
P.O. Box 2487
Kirkland, WA 98083

Institute for Personality and Ability Testing
P.O. Box 1188
Champaign, IL 61824-1188

Jastak Associates, Inc.
P.O. Box 3410
Wilmington, DE 19804-0250

Jossey-Bass
615 Montgomery Street
San Francisco, CA 94111

Mind Garden
P.O. Box 60669
Palo Alto, CA 94306

National Computer Systems
P.O. Box 1416
Minneapolis, MN 55440

National Rehabilitation Services
P.O. Box 1247
Gaylord, MI 49735

Oxford University Press
200 Madison Avenue
New York, New York 10016

Pro-Ed
8700 Shoal Creek Boulevard
Austin, TX 78757-6897

Psychological Assessment Resources, Inc.
P.O. Box 998
Odessa, FL 33556-0998

Psychological Corporation
555 Academic Court
San Antonio, TX 78204-2498

Reitan Neuropsychology Laboratories
2920 South 4th Avenue
Tucson, AZ 85713-4819

Riverside Publishing Company
3 O'Hare Towers
8420 Bryn Mawr Avenue
Chicago, IL 60631

Saul Rosenzweig
8029 Washington Avenue
St. Louis, MO 63114

Scholastic Tecting Service
480 Meyer Road
P.O. Box 1056
Bensenville, IL 60106-1617

Sigma Assessment Systems, Inc.
P.O. Box 610984
Port Huron, MI 48061-0984

SRA/London House
9701 W. Higgins Road
Rosemont, IL 60018

Slosson Educational Publications, Inc.
P.O. Box 280
East Aurora, NY 14052

Stoelting Company
620 Wheat Lane
Wood Dale, IL 60191

Charles C. Thomas
2600 South First Street
Springfield, IL 62794-9265

U.S. Employment Service
Western Assessment Research
and Development Center
140 East 300 South
Salt Lake City, UT 84111

U.S. Military Entrance
Processing Command
Testing Directorate
2500 Green Bay Road
North Chicago, IL 60064

University of Illinois Press
1325 South Oak Street
Champaign, IL 61820

Western Psychological Services
12031 Wilshire Boulevard
Los Angeles, CA 90025-1251

West Virginia Rehabilitation Center
509 Allen Hall
West Virginia University
Morgantown, WV 26506

Wilmington Institute
13315 Wilmington Drive
Dallas, TX 75234

Wonderlic Personnel Test, Inc.
1509 North Milwaukee Avenue
Libertyville, IL 60048-1380 ■

Apéndice C

Principales pruebas y sus editoriales

Pruebas individuales de inteligencia y conducta adaptativa

Adaptive Behavior Inventory for Children (Inventario de conducta adaptativa para niños); Psychological Corporation.

AAMR Adaptive Behavior Scale (Escala de conducta adaptativa AAMD); Pro-Ed.

AAMR Adaptive Behavior Scale –School Edition (Escala de conducta adaptativa AAMD –Edición escolar); Pro-Ed.

Balthazar Scales of Adaptive Behavior (Escalas de conducta adaptativa Balthazar); Consulting Psychologists Press.

Bayley Scales of Infant Development-II (Escalas de desarrollo infantil Bayley- II); Psychological Corporation.

The Blind Learning Aptitude Test (Prueba de aptitud de aprendizaje para invidentes); University of Illinois Press.

Bruininks-Oseretsky Test of Motor Proficiency (Prueba Bruininks-Oseretsky de destreza motora); American Guidance Service.

Columbia Mental Maturity Scale (Escala Columbia de madurez mental); Psychological Corporation.

Denver-2; Denver Developmental Materials.

Detroit Test of Learning Aptitude–3 (Pruebas Detroit de aptitud para el aprendizaje–3); Pro-Ed.

Developmental Indicators for the Assessment of Learning-Revised (Indicadores del desarrollo para la evaluación del aprendizaje–revisada); American Guidance Service.

Differential Ability Scales (Escalas de capacidad diferencial); Psychological Corporation.

Test del dibujo de la figura humana: sistema de calificación cuantitativa; Psychological Corporation.

Test de dibujo Goodenough-Harris; Psychological Corporation.

Hiskey-Nebraska Test of Learning Aptitude (Prueba Hiskey-Nebraska de aptitud para el aprendizaje); Marshal S. Hiskey.

Independent Living Behavior Checklist (Lista de verificación conductual de vida independiente); West Virginia Rehabilitation Center.

Batería de evaluación de Kaufman para niños; American Guidance Service.

Kaufman Brief Intelligence Test (Prueba breve de inteligencia de Kaufman); American Guidance Service.

Kaufman Adolescent and Adult Intelligence Test (Prueba Kaufman de inteligencia adolescente y adulta); American Guidance Service.

Leiter International Performance Scale-Revised (Escala internacional Leiter de desempeño–Revisada); Stoelting Company.

Escala McCarthy de aptitudes y psicomotricidad para niños; Psychological Corporation.

Miller Assessment for Preschoolers (Evaluación Miller para Preescolares); Psychological Corporation.

Ordinal Scales of psychological Development (Escalas ordinales de desarrollo psicológico); Uzgiriz y Hunt (1989).

Peabody Picture Vocabulary Test –Revised (Prueba Peabody de Vocabulario con Figura –Revisada); American Guidance Service.

Scales of Independent Behavior-R (Escalas de conducta independiente-R); DLM Teaching Resources.

Stanford-Binet: Cuarta edición; Riverside Publishing Co.

Test of Nonverbal Intelligence-3 (Prueba de inteligencia no verbal-3); Pro-Ed.

TMR School Competency Scales (Escalas de competencia escolar para DMT); Consulting Psychologists Press.

Vineland Adaptive Behavior Scales (Escalas Vineland de conducta adaptativa); American Guidance Service.

Escala Wechsler de inteligencia para adultos-III; Psychological Corporation.

Escala Wechsler de inteligencia para niños-III; Psychological Corporation.

Escala Wechsler de inteligencia para los niveles preescolar y primaria–revisada; Psychological Corporation.

Pruebas grupales de inteligencia

Cognitive Abilities Test (Prueba de capacidades cognitivas); Riverside Publishing Co.

Culture Fair Intelligence Test (Prueba de inteligencia culturalmente justa); Institute for Personality and Ability Testing.

Henmon-Nelson Tests of Mental Ability (Pruebas Henmon-Nelson de capacidad mental); Riverside Publishing Co.

Kuhlmann-Anderson Tests of Mental Ability (Pruebas Kuhlmann-Anderson de capacidad mental); Scholastic Testing Service.

Miller Analogies Test (Prueba Miller de analogías); Psychological Corporation.

Multidimensional Aptitude Battery (Batería multidimensional de aptitud); Sigma Assessment Systems.

Otis-Lennon School Ability Test (Prueba de capacidad escolar Otis-Lennon); Psychological Corporation.

Matrices progresivas de Raven; Psychological Corporation.

School and College Ability Tests –Series III (Pruebas de capacidad escolar y universitaria –Serie III); Educational Testing Service.

Shipley Institute of Living Scale (Escala del Shipley Institute of Living); Western Psychological Services.

Wonderlic Personnel Evaluation (Prueba Wonderlic de personal); Wonderlic Personnel Test.

Pruebas y baterías de aptitud

American College Testing Assessment Program (Programa de pruebas de evaluación universitaria estadounidense); American College Testing Program.

Armed Services Vocational Aptitude Battery (Batería de aptitudes vocacionales de las Fuerzas Armadas); U. S. Military Entrance Processing Command.

Test de aptitudes diferenciales; Psychological Corporation.

General Aptitude Test Battery (Batería de prueba de aptitud general); U. S. Employment Service.

Graduate Record Examinations (Exámenes de registro de posgrado); Educational Testing Services.

Scholastic Assessment Test (Pruebas de evaluación escolar); Educational Testing Service.

Pruebas grupales de rendimiento

California Achievement Tests (Pruebas de rendimiento de California); CTB/Macmillan/McGraw-Hill.

Comprehensive Tests of Basic Skills (Pruebas amplias de habilidades básicas); CTB/Macmillan/McGraw-Hill.

Iowa Tests of Basic Skills (Pruebas Iowa de habilidades básicas); Riverside Publishing Co.

Iowa Tests of Educational Development (Pruebas Iowa de desarrollo educativo); Riverside Publishing Co.

Metropolitan Achievement Tests (Pruebas de ejecución Metropolitan); Psychological Corporation.

Sequential Tests of Educational Progress, Series III (Pruebas secuenciales de progreso educativo, Serie III); Educational Testing Service.

SRA Achievement Series (Serie de rendimiento SRA); SRA/London House.

Stanford Achievement Series (Serie Stanford de rendimiento); SRA/London House.

Stanford Test of Academic Skills (Prueba Stanford de habilidades académicas); Psychological Corporation.

Test of Achievement and Proficiency (Pruebas de rendimiento y aprovechamiento); Riverside Publishing Co.

Tests of General Educational Development (GED) (Prueba de desarrollo educativo general; GED); GED Testing Service.

Pruebas individuales de rendimiento

Kaufman Test of Educational Achievement (Prueba Kaufman de rendimiento educativo); American Guidance Service.

Mini-Battery of Achievement (Minibatería de rendimiento); DLM Teaching Resources.

Peabody Individual Achievement Test-Revised (Prueba Peabody de ejecución individual –revisada); American Guidance Service.

Wechsler Individual Achievement Test (Prueba Wechsler de rendimiento individual); Psychological Corporation.

Wide Range Achievement Test-III (Prueba de rendimiento de amplio rango-III); Jastak Associates, Inc.

Woodcock-Johnson Psycho-Educational Battery-Revised (Batería psicoeducativa Woodcock-Johnson–revisada); Riverside Publishing Co.

Pruebas psicomotoras y de destreza

Crawford Small Parts Dexterity Test (Prueba de destreza de partes pequeñas de Crawford); Psychological Corporation.

Hand-Tool Dexterity Test (Prueba de destreza mano-herramienta); Psychological Corporation.

Test no verbal de la Universidad de Purdue; SRA/London House.

Stromberg Dexterity Test (Prueba Stromberg de destreza); Psychological Corporation.

Pruebas para trabajos de oficina

Clerical Abilities Battery (Batería de capacidades para trabajos de oficina); Psychological Corporation.

General Clerical Test (Prueba general de trabajos de oficina); Psychological Corporation.

Minnesota Clerical Test (Prueba Minnesota de trabajos de oficina); Psychological Corporation.

SRA Clerical Aptitudes (Aptitudes para trabajos de oficina SRA); SRA/London House.

Prueba de aptitudes mecánicas

Bennett Mechanical Comprehension Test (Prueba Bennett de comprensión mecánica); Psychological Corporation.

Minnesota Spatial Relations Test (Prueba Minnesota de relaciones espaciales); American Guidance Service.

Revised Minnesota Paper Form Board Test (Prueba Minnesota revisada de tablero con formas de papel); Psychological Corporation.

SRA Mechanical Aptitudes (Aptitudes mecánicas SRA); SRA/London House.

Inventarios de intereses

Campbell Interest and Skill Survey (Estudio de intereses y habilidades de Campbell); National Computer Systems.

Jackson Vocational Interest Survey (Encuesta Jackson de interés Vocacional); Sigma Assessment Systems.

Kuder General Interest Sulvey (Estudio de intereses generales de Kuder); SRA/London House.

Kuder Occupational Interest Survey (Estudio de intereses laborales) SRA/London House.

Kuder Preference Record (Registro Kuder de preferencias); SRA/London House.

Self–Directed Search (Búsqueda autodirigida); Psychological Assessment Resources, Inc.

Strong Interest Inventory (Inventario Strong de intereses); Consulting Psychologists Press.

Pruebas neuropsicológicas

Test Guestáltico Visomotor; Western Psychological Services.

Benton Revised Retention Test (Prueba Benton de retención visual, revisada); Psychological Corporation.

Finger Localization Test (Prueba de localización de dedos); Oxford University Press.

Halstead-Reitan Neuropsychological Test Battery (Batería de pruebas neuropsicológicas Halstead-Reitan); Reitan Neuropsychology Laboratories.

Luria-Nebraska Neuropsychological Battery (Diagnóstico Neuropsicológico Luria-Nebraska); Western Psychological Services.

Test de laberintos Porteus; Psychological Corporation.

Serial Digit Learning Test (Prueba de aprendizaje de series de dígitos); Oxford University Press.

Symbol Digit Learning Test (Prueba de modalidades de símbolos y dígitos); Western Psychological Services.

Three-Dimensional Block Construction Test (Prueba de construcción con cubos tridimensionales); Oxford University Press.

Wechsler memory Scale–Revised (Escala Wechsler de memoria-revisada); Psychological Corporation.

Wisconsin Card Sorting Test (Prueba Wisconsin de clasificación de tarjetas); Psychological Assessment Resources.

Pruebas proyectivas de personalidad

Test de apercepción infantil; C.P.S. Inc.

Test del dibujo de la figura humana; Charles C. Thomas.

Holtzman Inkblot Test (Prueba Holtzman de manchas de tinta); Psychological Corporation

Rorschach; Hogrefe & Huber Publishers

Rosenzweig Picture-Frustration Study (P-F Study) (Estudio Rosenzweig de frustración por ilustraciones; Estudio P-F); Saul Rosenzweig.

Rotter Incomplete Sentences Blank (Formulario de oraciones incompletas de Rotter); Psychological Corporation.

Técnica de apercepción para edades avanzadas; C.P.S., Inc.

Test de apercepción temática (TAT); Harvard University Press.

Washington University Sentence Completion (Prueba de frases incompletas de la Universidad de Washington); Jossey-Bass.

Inventarios de autoinforme de la personalidad

Beck Depression Inventory (Inventario de depresión Beck, revisado); Psychological Corporation.

Inventario psicológico de California; Consulting Psychologists Press.

Comrey Personality Scales (Escalas de personalidad Comrey), EdITS; Educational and Industrial Testing Service.

Edwards Personal Preference Schedule (Inventario de preferencias personales de Edwards); Psychological Corporation.

Eysenck Personality Questionnaire (Cuestionario Eysenck de personalidad), EdITS; Educational and Industrial Testing Service.

Jenkins Activity Survey (Inventario Jenkins de actividad); Psychological Corporation.

Millon Clinical Multiaxial Inventory-3 (Inventario clínico multiaxial de Millon–3); National Computer Systems.

Inventario multifásico de la personalidad de Minnesota–2; National Computer Systems.

Inventario Tipológico de Myers-Briggs; Consulting Psychologists Press.

NEO Five Factor Inventory (Inventario NEO de cinco factores); Sigma Assessment Systems.

NEO Personality Inventory-Revised (Inventario NEO de personalidad); Sigma Assessment Systems.

Personality Inventory for Children (Inventario de personalidad para niños); Western Psychological Services.

Personality Research Form (Forma de investigación de la personalidad); Sigma Assessment Systems.

Cuestionario de 16 factores de la personalidad (16FP®); Institute for Personality and Ability Testing.

Inventario de ansiedad: rasgo-estado; Mind Garden.

Survey of Work Styles (Estudio de estilos de trabajo); Sigma Assessment Systems.

Pruebas forenses

Custody Quotient™; Wilmington Institute.

Parent-Child Relationship Inventory (Inventario de relaciones padres-hijo); Western Psychological Services

Rogers Criminal Responsibility Assessment Scales (Escalas Rogers de evaluación de la responsabilidad criminal); Psychological Assessment Resources, Inc. ■

Equivalentes de rangos percentiles en puntuación estándar y estandarizada en una distribución normal

APÉNDICE D

Esta tabla lista las equivalencias entre rangos percentiles y otros cuatro tipos de puntuaciones: puntuaciones *z* (media de 0, DE de 1.00); CI de desviación (media de 100, DE de 15); puntuaciones *T* (media de 50, DE de 10) y puntuaciones del tipo GRE (media de 500, DE de 100. La aplicación de la tabla supone que la distribución de calificaciones en una prueba o variable es normal.

Se ilustra con dos ejemplos la manera en que puede utilizarse este apéndice. Supóngase que se tiene el deseo de conocer el CI del WAIS-R que es equivalente a un rango percentil de 97. Si se lee a lo largo de la fila que comienza con RP 97, se descubrirá que el CI equivalente es de 128. Supóngase que la intención es conocer el rango percentil equivalente a una puntuación GRE de 675. En la columna al extremo derecho se localiza una calificación de 675 y se lee a lo largo de la fila hasta la columna al extremo izquierdo para descubrir que el rango percentil equivalente es 96.

	z	CI de desviación	Puntuación *T*	Puntuación tipo GRE
Media	0.00	100	50	500
Des. est.	1.00	15	10	100
RP 99	2.33	135	73	733
98	2.05	131	71	705
97	1.88	128	69	688
96	1.75	126	68	675
95	1.64	125	66	664
94	1.55	123	66	655
93	1.48	122	65	648
92	1.41	121	64	641
91	1.34	120	63	634
90	1.28	119	63	628
89	1.22	118	62	622
88	1.18	118	62	618
87	1.13	117	61	613
86	1.08	116	61	608
85	1.04	116	60	604
84	0.99	115	60	599
83	0.95	114	60	595
82	0.91	114	59	591
81	0.88	113	59	588
80	0.84	113	58	584
79	0.80	112	58	580
78	0.77	112	58	577
77	0.74	111	57	574
76	0.71	111	57	571
75	0.67	110	57	567
74	0.64	110	56	564
73	0.61	110	56	561
72	0.58	109	56	558
71	0.55	108	56	555
70	0.52	108	55	552
69	0.49	107	55	549
68	0.47	107	55	547
67	0.44	107	54	544
66	0.41	106	54	541

(continuación)

	z	CI de desviación	Puntuación *T*	Puntuación tipo GRE
65	0.39	106	54	539
64	0.36	105	54	536
63	0.33	105	53	533
62	0.31	105	53	531
61	0.28	104	53	528
60	0.25	104	53	525
59	0.23	104	52	523
58	0.20	103	52	520
57	0.18	103	52	518
56	0.15	102	52	515
55	0.12	102	51	512
54	0.10	102	51	510
53	0.07	101	51	507
52	0.05	101	51	505
51	0.03	100	50	503
50	0.00	100	50	500
49	–0.03	100	50	497
48	–0.05	99	49	495
47	–0.07	99	49	493
46	–0.10	98	49	490
45	–0.12	98	49	488
44	–0.15	98	48	485
43	–0.18	97	48	482
42	–0.20	97	48	480
41	–0.23	96	48	477
40	–0.25	96	47	475
39	–0.28	96	47	472
38	–0.31	95	47	469
37	–0.33	95	47	467
36	–0.36	95	46	464
35	–0.39	94	46	461
34	–0.41	94	46	459
33	–0.44	93	46	456
32	–0.47	93	45	453
31	–0.49	93	45	451
30	–0.52	92	45	448

(continuación)

	z	CI de desviación	Puntuación T	Puntuación tipo GRE
29	−0.55	92	44	445
28	−0.58	91	44	442
27	−0.61	90	44	439
26	−0.64	90	44	436
25	−0.67	90	43	433
24	−0.71	89	43	429
23	−0.74	89	43	426
22	−0.77	88	42	423
21	−0.80	88	42	420
20	−0.84	87	42	416
19	−0.88	87	41	412
18	−0.91	86	41	409
17	−0.95	86	40	405
16	−0.99	85	40	401
15	−1.04	84	40	396
14	−1.08	84	39	392
13	−1.13	83	39	387
12	−1.18	82	38	382
11	−1.22	82	38	378
10	−1.28	81	37	372
9	−1.34	80	37	366
8	−1.41	79	36	359
7	−1.48	78	35	352
6	−1.55	77	34	345
5	−1.64	75	34	336
4	−1.75	74	32	325
3	−1.88	72	31	312
2	−2.05	69	29	295
1	−2.33	65	27	267

Glosario

Acomodación. En la teoría de Piaget, la adaptación de un esquema que no ha tenido éxito, a fin de que funcione.

Afasia. Cualquier desviación en el desempeño en el lenguaje, producida por daño cerebral.

Agnosia visual. Dificultad en el reconocimiento de dibujos, objetos o rostros producida por daño cerebral.

Alegato de demencia. En casos judiciales, defensa basada en la referencia a la demencia legal como la determinan el Código Penal Modelo u otros estatutos legales.

Amenaza del estereotipo. La amenaza de confirmar, como característica propia, un estereotipo negativo acerca del propio grupo.

Análisis de puestos. Proceso para definir un puesto en términos de las conductas necesarias para desempeñarlo; incluye descripción del puesto (características físicas del mismo) y especificación del mismo (características personales necesarias).

Análisis factorial. Familia de procedimientos estadísticos que utilizan los investigadores para resumir las relaciones entre variables

que se correlacionan de modos sumamente complejos; la meta del análisis factorial consiste en derivar un conjunto parsimonioso de factores derivados.

Ansiedad de prueba. Una constelación de respuestas fenomenológicas, fisiológicas y conductuales que acompañan la preocupación acerca del posible fracaso en una prueba.

Ansiedad estado. Sentimientos transitorios de temor o preocupación que experimentan la mayoría de las personas en ciertas ocasiones.

Ansiedad rasgo. Tendencia relativamente estable de un individuo a responder de manera ansiosa a situaciones estresantes.

Apraxia. Variedad de trastornos caracterizados por una falla en la dirección o ejecución de actos motores complejos.

Asignación. En evaluación con pruebas, la distribución de las personas en diferentes programas apropiados para sus necesidades o habilidades.

Asimilación. En la teoría de Piaget, la aplicación de un esquema a un objeto, persona o acontecimiento.

Autista sabio. Individuo con retraso mental que tiene un talento sumamente desarrollado en una sola área, como arte, cálculo rápido, memoria o música.

Autoeficacia. En la teoría de Bandura, juicio personal acerca de qué tan bien puede uno ejecutar los cursos de acción requeridos para manejar situaciones futuras.

Automonitoreo. Enfoque terapéutico en el que el cliente elige las metas y participa de manera activa en la supervisión, graficación y registro del progreso hacia los fines de la terapia.

Ayuda. En exámenes psicológicos, intento por mejorar las calificaciones de una prueba al proporcionarle al individuo evaluado práctica adicional en materiales similares, revisión de conceptos fundamentales que es probable que se cubran en la prueba y consejo acerca de las estrategias óptimas para la presentación de la misma.

Carga factorial. En el análisis factorial, la correlación entre una prueba individual y un solo factor.

Centro de evaluación. Enfoque para evaluar el talento gerencial que consiste en múltiples técnicas de simulación, incluyendo presentaciones de grupo, ejercicios de solución de problemas, ejercicios de discusión grupal, entrevistas y técnicas de bandeja de entrada.

Cerebelo. Parte del metencéfalo responsable de ayudar a coordinar tono muscular, postura y movimientos hábiles.

Certificación. Examen para determinar que una persona tiene cuando menos un mínimo de destreza en alguna disciplina o actividad.

Clasificación de expertos. Un método de escala que depende de los juicios de expertos para determinar las clasificaciones de los componentes individuales.

Clasificación. En exámenes psicológicos, el proceso de utilizar pruebas para asignar a una persona dentro de una categoría más que de otra.

Coeficiente alfa. Índice de confiabilidad que puede considerarse como la media de todos los coeficientes posibles por mitades, corregido según la fórmula de Spearman-Brown.

Coeficiente de confiabilidad. Proporción de varianza verdadera de la puntuación con respecto a la varianza total de las calificaciones de la prueba.

Coeficiente de correlación. Índice numérico del grado de relación lineal entre dos conjuntos de calificaciones; los coeficientes de correlación pueden variar entre -1.00 y +1.00.

Coeficiente de personalidad. Término utilizado para referirse al descubrimiento de que la validez predictiva de las escalas de personalidad supera en pocas ocasiones al .30.

Coeficiente de validez. Correlación entre la prueba y el criterio (r_{xy}).

Competencia para someterse a juicio. Determinación del juez que preside la sala acerca de que un acusado **no** tiene defecto, enfermedad o padecimiento mental que le incapacite para comprender el proceso o para ayudar en su defensa.

Comprensión verbal. Primer factor en el WISC-III que incluye Información, Semejanzas, Vocabulario y Comprensión.

Conducta no verbal. Formas sutiles de comunicación humana contenidas en la mirada, ademanes, lenguaje corporal, tono de voz y expresión verbal.

Confiabilidad de formas alternas. Método de confiabilidad en el que se aplican formas alternas de una misma prueba a un grupo de individuos heterogéneos y representativos; después se correlacionan las calificaciones de cada forma.

Confiabilidad intercalificadores. Para pruebas que implican calificación según un juicio, grado común de concordancia entre calificadores.

Confiabilidad por mitades. Forma de confiabilidad en la que las calificaciones de ambas mitades de una prueba se correlacionan entre sí (p. ej., reactivos pares contra nones); la correlación se ajusta entonces en cuanto a la extensión de la prueba.

Confiabilidad *test-retest*. Forma de confiabilidad en la que se aplica la misma prueba dos veces al mismo grupo de individuos heterogéneos y representativos; después se correlacionan las calificaciones de ambas sesiones.

Confiabilidad. Atributo de consistencia en la medición.

Conmoción cerebral. Alteración transitoria de la conciencia debida a un golpe en la cabeza; puede ir seguida por amnesia temporal, mareo, náuseas, pulso débil y respiración lenta; sin embargo, no existe daño cerebral orgánico demostrable.

Consentimiento informado. En pruebas, el principio de informar a la persona evaluada o a sus representantes, en un lenguaje que puedan comprender, los propósitos de las pruebas y sus posibles consecuencias.

Conservación. En la teoría de Piaget, conciencia de que las cantidades físicas no cambian cuando se les altera superficialmente.

Constancia del CI. En las pruebas Wechsler, suposición axiomática de que el CI debe permanecer constante a medida que se envejece normalmente, aun cuando la capacidad intelectual bruta pueda cambiar o declinar.

Constructo. Cualidad o rasgo teórico e intangible en el cual difieren los individuos.

Contaminación del criterio. Fuente de error en la validación de una prueba cuando el criterio se "contamina" debido a su semejanza artificial con la misma, de modo que la prueba y el criterio contienen reactivos casi idénticos. También es una forma de valorar el error en el cual una medida de criterio incluye factores que no son parte demostrable del trabajo, por ejemplo, la aparente clasificación cuando éste no es trabajo-relacionado.

Corrección de conjeturas. En las pruebas grupales, la práctica de revisar la calificación final del individuo cuando existe aparente conjeturación.

Corteza cerebral. Capa externa del cerebro, que es la fuente de los niveles superiores de procesamiento sensorial, motor y cognitivo.

Cuadro de expectativas. Un cuadro que representa la relación establecida entre las calificaciones de una prueba y los resultados esperados en una relacionada.

Cuerpo calloso. La principal comisura que sirve para integrar las funciones de ambos hemisferios cerebrales.

Culpable pero mentalmente enfermo. Veredicto permitido en algunos estados en el que la intención es que el acusado comience su condena en un hospital psiquiátrico.

Cuotas. En pruebas de selección, la postura ética de que los candidatos más calificados dentro de subgrupos definibles deben seleccionarse en proporción a su representación dentro de la población.

Curva reactivo-característica. Representación gráfica de la relación entre la probabilidad de una respuesta correcta y la posición de la persona evaluada en el rasgo subyacente medido por la prueba.

Daños personales. En las demandas por daños personales, los abogados pueden contratar a psicólogos para que testifiquen con relación a las consecuencias crónicas del estrés traumático o del daño cerebral adquirido.

Datos biográficos. Datos autobiográficos objetivos o calificables: reconocidos como complemento válido para la selección de personal.

Deber de advertir. Derivado del caso Tarasoff, se refiere a la responsabilidad del clínico de comunicar cualquier amenaza grave a la víctima potencial, organismos encargados de velar por el cumplimiento de la ley o ambos.

Decreto sobre los estadounidenses con discapacidades. Ley aprobada por el Congreso de EUA en 1990 que prohíbe la discriminación contra individuos calificados que tienen discapacidades.

Definición funcionalista de validez. Perspectiva de que una prueba es válida si sirve para el propósito para el que se le utiliza.

Definición operacional. Definición de un concepto en términos de la manera en que se mide; como inteligencia es "lo que miden las pruebas".

Definición real. Una definición que intenta decirnos la verdadera naturaleza de las cosas que se definen.

Desviación estándar. Índice estadístico que refleja el grado de dispersión en un grupo de calificaciones; la raíz cuadrada de la varianza.

Detección. Uso de pruebas o procedimientos rápidos y simples para identificar a las personas que podrían tener características o necesidades especiales.

Diagnóstico. Determinación de la naturaleza y fuente de la conducta anormal de una persona y clasificación del patrón conductual dentro de un sistema diagnóstico aceptado.

Diferencial semántico. Técnica de calificación en la que el individuo utiliza un continuo de siete puntos para calificar un concepto o varios adjetivos bipolares como bueno-malo, fuerte, débil, activo-pasivo.

Dilema moral. Historia breve que implica una elección moral difícil, como robar para prolongar la vida de otra persona; se le utiliza en el estudio del razonamiento moral.

Discapacidad para el aprendizaje. Concepto indistinto que por lo común se refiere a una discrepancia grave entre la capacidad general y el rendimiento individual que no puede explicarse por discapacidad sensorial/motora, retraso mental, problemas emocionales o privación cultural.

Diseño longitudinal. Diseño de investigación en el que se examina a los mismos individuos en diversos puntos en el tiempo.

Diseño transecuencial. Diseño de investigación que combina los métodos transversal y longitudinal.

Diseño transversal. Diseño de investigación en el que se examina a individuos de diferentes edades en algún punto en el tiempo.

Dispraxia de construcción. Alteración de la capacidad para manejar relaciones espaciales en un esquema, ya sea de 2 o 3 dimensiones.

Distribución de frecuencias. Método para resumir los datos o calificaciones de una prueba al especificar un pequeño número de intervalos de clase que por lo general son de igual tamaño y después registrar cuántas calificaciones caen dentro de cada intervalo.

Distribución normal. Distribución simétrica de frecuencias, definida en términos matemáticos y en forma de campana.

Economía de fichas. Enfoque conductual en el que se recompensan muchas diferentes formas de conducta prosocial a través de fichas que pueden intercambiarse después por recompensas materiales o por privilegios.

Ecuación de regresión. Ecuación que describe la línea recta que más se ajusta para estimar el criterio a partir de la prueba; la línea que más se ajusta es aquella que reduce al mínimo las desviaciones al cuadrado con respecto a la misma.

Efecto alcohólico fetal. Versión leve del síndrome alcohólico fetal en el que no se observan anormalidades físicas, pero se notan problemas conductuales como dificultades en atención.

Efecto de halo. Tendencia a calificar a un empleado como alto o bajo en todas las dimensiones debido a una impresión global.

Efecto del Lago Wobegon. Observación de que casi todos los estados de EUA afirman que las calificaciones promedio de rendimiento para sus sistemas escolares superan al percentil 50.

Ejercicio situacional. Procedimiento de evaluación en el que se pide al futuro empleado que desempeñe el trabajo bajo circunstancias que son sumamente similares al ambiente laboral anticipado.

Ejes oblicuos. En análisis factorial, la suposición de que los factores se correlacionan entre sí; es decir, no están en ángulos rectos.

Ejes ortogonales. En análisis factorial, la suposición de que los factores están en ángulo recto entre sí, lo cual significa que permanecen intercorrelacionados.

Ello. En la teoría psicoanalítica, parte inconsciente de la personalidad donde residen todas las necesidades instintivas, como aquéllas de comida, agua, gratificación sexual y evitación del dolor.

Enfermedad de Alzheimer. Trastorno neurológico degenerativo; en las primeras etapas el síntoma más notable es la pérdida de memoria.

Enfoque adaptado al criterio. Enfoque de desarrollo de pruebas en el que los reactivos se asignan a una escala particular si, y sólo si, discriminan entre un grupo criterio bien definido y un control relacionado.

Equidad de una prueba. El grado al cual las consecuencias sociales del uso de pruebas se consideran justas para subgrupos relacionados; la equidad de una prueba es una cuestión de valores sociales y es pertinente en especial cuando las pruebas se utilizan para tomar decisiones de selección.

Equilibración. En la teoría de Piaget, proceso completo de asimilación, acomodación y equilibrio.

Error administrativo de calificación. En exámenes psicológicos, un error en la calificación de una prueba relacionado con la mecánica de calificación; como sumar de manera incorrecta las calificaciones o consultar la tabla de conversión errónea.

Error de medición. Toda otra calificación diferente a la verdadera que integra la calificación total obtenida por la persona evaluada.

Error estándar de estimación. SE_{est} es el margen de error que puede esperarse en la puntuación de criterio predicha; el error de estimación se obtiene de la siguiente fórmula:

$$SE_{est} = SD_y\sqrt{1 - r_{xy}^2}$$

Error estándar de la diferencia. Índice estadístico que puede ayudar al usuario de una prueba a determinar si, en el caso de un individuo examinado, la diferencia entre las calificaciones de dos pruebas o subpruebas es significativa.

Error estándar de medición. Índice del error de medición que indica el grado al que las calificaciones de una persona examinada podrían variar en varias pruebas paralelas.

Error sistemático de medición. Tipo de error de medición que surge cuando, sin que lo sepa quien desarrolló la prueba, el instrumento mide alguna otra cosa diferente al rasgo al cual se dirige.

Escala C. Variante de la escala de estaninas con 11 unidades.

Escala de clasificación con base conductual. Escala de clasificación referida a criterio.

Escala de elección forzosa. Escala de evaluación del desempeño diseñada para eliminar el sesgo y subjetividad en las calificaciones del supervisor al forzar una elección entre opciones iguales en cuanto a deseabilidad social.

Escala de estándar mixto. Enfoque complejo para la valoración del desempeño diseñado para reducir al mínimo los errores de calificación en la evaluación del desempeño; los reactivos para las dimensiones de desempeño se ordenan de manera aleatoria en la escala.

Escala de estaninas. Una escala en la que todas las puntuaciones naturales se convierten a un sistema de calificaciones de un solo dígito que va de 1 a 9.

Escala de estenes. Escala de 10 unidades con cinco unidades por encima y cinco unidades por debajo de la media.

Escala de intervalo. Escala de medición que proporciona información acerca de la clasificación por rangos y fortaleza relativa de los mismos; basada en la suposición de unidades o intervalos de igual tamaño para la escala subyacente.

Escala de observación conductual. Una variación sobre la técnica BARS, que utiliza un continuo desde "casi nunca" hasta "casi siempre" para medir la frecuencia con la que un empleado lleva a cabo tareas específicas en cada dimensión conductual.

Escala gráfica de calificación. Escala que incluye etiquetas de rasgo, definiciones breves de dichas etiquetas y un continuo para calificación.

Escala homogénea. Escala en la que los reactivos individuales tienden a medir la misma cosa; la homogeneidad se estima a través de correlaciones reactivo-total.

Escala Likert. Escala que le presenta a la persona examinada cinco respuestas ordenadas en un continuo de acuerdo/desacuerdo o aprobación/desaprobación.

Escala nominal. Escala de medición en la que las categorías son arbitrarias y no designan "más" o "menos" de ninguna cosa; el nivel más sencillo y bajo de medición.

Escala ordinal. Escala de medición que permite una clasificación; las escalas ordinales no proporcionan información acerca de la fortaleza relativa de la clasificación.

Escala radio. Escala de medida que produce unidades o intervalos de igual tamaño y posee un punto cero conceptualmente significativo; el nivel más alto de medida.

Esquema. En la teoría de Piaget, patrón organizado de conducta o estructura mental bien definida que conduce al conocimiento de cómo hacer alguna cosa.

Estado mental al momento de la ofensa (EMO). El estado mental de un acusado al momento de cometer la ofensa es pertinente en defensas especiales como la defensa de insania; los psicólogos y psiquiatras pueden ofrecer opiniones acerca del EMO de un acusado.

Estándar de cuidado. Las normas de cuidado que son comunes, acostumbradas o razonables.

Evaluación conductual. Variedad de técnicas que se concentran en la conducta en sí, más que en los rasgos subyacentes, causas hipotéticas o supuestas dimensiones de personalidad.

Evaluación psicológica asistida por computadora, CAPA. Se refiere al rango completo de aplicaciones computacionales en la evaluación psicológica e incluye aplicación, calificación, redacción del informe y aplicación individualizada de la prueba.

Evaluación. Valoración o estimación del nivel o magnitud de algún atributo de una persona; las pruebas son sólo una pequeña parte de la evaluación, que también incluye ob-

servaciones, entrevistas, escalas de calificación y listas de verificación.

Expresión religiosa extrínseca. Uso de la religión para metas externas como seguridad, estatus y amistad.

Expresión religiosa intrínseca. Utilización de la religión para metas internas como encontrar significado y dirección en la vida.

Extraversión. Disposición de personalidad sociable, comunicativa y de búsqueda de emociones.

Factor específico. Según Spearman, un factor de la inteligencia específico de una prueba individual.

Factor general. Según Spearman, el factor general de inteligencia que debe existir para explicar las correlaciones observadas entre un gran número de pruebas.

Factor. Constructo o variable subyacente que ayuda a explicar las correlaciones entre varias pruebas o medidas.

Falacia de la estandarización. Perspectiva engañosa de que una prueba estandarizada en una población es injusta *ipso facto* cuando se le utiliza con cualquier otra población.

Falso negativo. En la teoría de decisión, un individuo de quien se ha pronosticado de manera incorrecta que fracasará en el criterio.

Falso positivo. En la teoría de decisión, un individuo de quien se ha pronosticado de manera incorrecta que tendrá éxito en el criterio.

Formación reticular. Red de cuerpos y fibras neuronales ascendentes y descendentes que gobiernan la excitación general o la conciencia.

Fórmula Kuder-Richardson 20. Índice de la confiabilidad que es pertinente para el caso especial donde cada reactivo de una prueba se califica con 0 o 1 (p. ej., correcto o incorrecto).

Fórmula Spearman-Brown. Formula para ajustar las correlaciones por mitades, de modo que reflejen la extensión total de una escala.

Frustración. En el sistema de Rosenzweig, estado que ocurre cuando un organismo se enfrenta a un obstáculo u obstrucción en la búsqueda de satisfacer una necesidad.

Funciones ejecutivas. Funciones cerebrales que incluyen análisis lógico, formación de conceptos, razonamiento, planificación y flexibilidad de pensamiento.

Ganglios basales. Conjunto de núcleos del prosencéfalo que se conectan con la corteza cerebral en la parte superior y el tálamo en la parte inferior, los ganglios basales participan en el control del movimiento.

Glándula pineal. Estructura del tamaño de un guisante que se encuentra en el centro del cerebro; secreta la hormona melatonina en un ritmo biológico cíclico, pero sus funciones no se han comprendido del todo.

Grupo normativo. Muestra de personas evaluadas que son representativas de la población hacia la cual se dirige la prueba.

Habilidades mentales primarias. Siete factores grupales de la inteligencia propuestos por Thurstone.

Hipocampo. Parte de un circuito de memoria complejo, indefinido que consolida las experiencias nuevas en recuerdos a largo plazo.

Hipotálamo. Pequeña estructura al centro del cerebro que ayuda a gobernar la conducta motivada y la regulación corporal: alimentación, conducta sexual, sueño, regulación de temperatura, conducta emocional y movimiento.

Hipótesis léxica fundamental. En la teoría de la personalidad, concepto de que los términos para rasgo han sobrevivido en el lenguaje debido a que trasmiten información importante sobre nuestros tratos con otras personas.

Hipótesis proyectiva. Suposición de que las interpretaciones personales de los estímulos ambiguos deben necesariamente reflejar las necesidades, motivos y conflictos inconscientes de la persona examinada.

Histograma. Método para resumir datos o calificaciones de una prueba en forma gráfica; éste contiene la misma información que una distribución de frecuencias.

Impacto adverso. En la contratación de personal, se dice que existe un impacto adverso si un grupo tiene una tasa de selección menor a cuatro quintas partes de la tasa del grupo con el nivel de selección más alto (*Uniform Guidelines on Employee Selection*, 1978).

Índice de confiabilidad del reactivo. $s_i r_{iT}$, el producto de la consistencia interna de un reactivo de una prueba como se indexa por la correlación con la calificación total (r_{iT}) y su variabilidad como se indexa por la desviación estándar (s_i).

Índice de deterioro intelectual. En la Escala del *Shipley Institute of Living*, índice basado en la discrepancia entre la capacidad verbal y de abstracciones que tiene el propósito de estimar los efectos de la discapacidad cerebral orgánica.

Índice de dificultad del reactivo. Para un solo reactivo de una prueba, proporción de individuos examinados en una gran muestra del piloteo que responden correctamente ese reactivo.

Índice de discriminación del reactivo. Índice estadístico de la eficacia con la que un reactivo discrimina entre personas que obtienen calificaciones altas y bajas en la prueba completa.

Índice de hereditabilidad. Estimado de aquella parte de la varianza total en un rasgo dado que se debe a factores genéticos; el índice puede variar desde 0.0 a 1.0

Índice de la validez del reactivo. $s_i r_{iC}$, consiste del producto de la desviación estándar de un reactivo de prueba (s_i) y la correlación punto-biserial con el criterio, r_{iC}.

Individualismo calificado. En pruebas de selección, la postura ética de que no deben utilizarse la edad, sexo, raza y otras características demográficas, incluso si el conocimiento de dichos factores pudiera mejorar la validez de la selección.

Individualismo no calificado. En pruebas de selección, postura ética en la que, sin excepción, debe seleccionarse para el puesto, admisiones u otros privilegios a los candidatos mejor calificados.

Inteligencia componencial. En la teoría de Sternberg, mecanismos mentales internos responsables de la conducta inteligente.

Inteligencia contextual. En la teoría de Sternberg, actividad mental implicada en la adaptación propositiva, moldeamiento y selección de ambientes reales pertinentes a la propia vida.

Inteligencia cristalizada. En la teoría de Cattell y Horn, aquello que ya se ha aprendido por la inversión de la inteligencia fluida en ambientes culturales (p. ej. aprender álgebra en la escuela).

Inteligencia experiencial. En la teoría de Sternberg, capacidad para enfrentarse de manera eficaz con tareas novedosas.

Inteligencia fluida. En la teoría de Cattell y Horn, una forma de eficiencia mental principalmente no verbal y que tiene una influencia cultural relativamente reducida.

Inteligencia social. Capacidad para comprender a otras personas y relacionarse de manera eficaz con ellas.

Inteligencia. Según los expertos, 1) la capacidad para aprender de la experiencia y 2) la capacidad para adaptarse al propio ambiente.

Interpretación de una prueba basada en sistemas, CBTI. Se refiere a la interpretación de una prueba y redacción del informe hechos por sistemas, que es uno de los principales componentes de la evaluación psicológica asistida por computadora (CAPA).

Introversión. Disposición de personalidad tranquila, reservada, de "ratón de biblioteca".

Inventario de estudio del temor. Instrumento de evaluación conductual que requiere que

aquellos que lo responden indiquen la presencia e intensidad de sus temores en relación con diversos estímulos, por lo común en una escala tipo Likert de 5 o 7 puntos.

Inventario de intereses. Prueba que mide la preferencia por ciertas actividades o temas y, por tanto, ayuda a determinar la elección ocupacional.

Juicio actuarial. Tipo de juicio automatizado en el que se utiliza una fórmula derivada empíricamente para diagnosticar o predecir la conducta.

Juicio clínico. Tipo de juicio en el que la persona que toma la decisión procesa la información dentro de su cabeza para diagnosticar o predecir la conducta.

Ley Pública 93-112. Una "declaración de derechos" para las personas con discapacidad que proscribe la discriminación con base en el impedimento.

Ley Pública 94-142. Decreto de Educación para Todos los Niños con Discapacidades (Ley Pública 94-142), que ordena que los escolares con impedimentos reciban evaluación apropiada y oportunidades educativas.

Ley Pública 99-457. Legislación que requiere que los estados proporcionen educación pública gratuita y apropiada para los niños discapacitados de 3 a 5 años de edad.

Libertad de la distractibilidad. Tercer factor del WISC-III que consiste de Aritmética y Retención de Dígitos.

Lista de verificación de incidentes críticos. Forma de evaluación del desempeño basada en episodios reales de conducta deseable e indeseable en el trabajo.

Lóbulo frontal. Parte de la corteza cerebral al frente del cerebro que se requiere para la programación, regulación y verificación de las funciones ejecutivas y de la ejecución motora.

Lóbulo límbico. Grupo de estructuras subcorticales responsables de la elaboración de la emoción y el control de la actividad visceral.

Lóbulo occipital. Parte de la corteza cerebral en la parte trasera del cerebro que contiene los centros de la visión.

Lóbulo parietal. Parte de la corteza cerebral que media la integración espacial y la conciencia sensorial de lo que sucede en la superficie del cuerpo.

Lóbulo temporal. Parte de la corteza cerebral implicada en el procesamiento de las sensaciones auditivas, almacenamiento de memoria a largo plazo y modulación de impulsos biológicos como agresión, temor y sexualidad.

***Locus* de control.** Constructo que se refiere a las percepciones que tienen las personas acerca de la fuente de aquello que les sucede (p. ej., interno contra externo).

Mal de Parkinson. Patología cerebral degenerativa caracterizada por tres tipos de trastornos motores: movimientos involuntarios, que incluyen temblor; pobreza y lentitud de movimientos, sin llegar a parálisis; y cambios en postura y tono muscular.

Manual del usuario. En pruebas, el manual que proporciona las instrucciones para aplicación y que también provee las pautas para interpretación de la prueba.

Manual técnico. En pruebas, el manual que resume los datos técnicos de un nuevo instrumento.

Matriz de correlación. Tabla completa de intercorrelaciones entre todas las variables, la cual es el punto inicial del análisis factorial.

Matriz factorial. Tabla de correlaciones entre variables y factores; éstas se denominan cargas factoriales.

Matriz multirrasgo-multimétodo. Diseño de investigación para evaluar validez convergente y de discriminación que requiere de la evaluación de dos o más rasgos a través de dos o más métodos.

Mecanismos de defensa. Estrategias mentales inconscientes disponibles para el yo cuando enfrenta las demandas conflictivas del ello, superyó y la realidad externa.

Media. Promedio aritmético de un grupo de calificaciones.

Mediana. Puntuación a la mitad de todas las calificaciones clasificadas.

Médula ósea. Parte del metencéfalo que ayuda a mediar los reflejos de tragar, vomitar, respirar; control de la presión sanguínea, respiración y, en parte, frecuencia cardiaca.

Mesencéfalo. Porción media del cerebro que incluye nervios craneales y estaciones de relevo para la visión y audición.

Metencéfalo. Estructuras cerebrales más bajas y de organización más simple.

Método de concordancia empírica. Método de desarrollo de escalas en el que los reactivos de las pruebas se seleccionan por completo con base en qué tan bien diferencian a un grupo criterio de una muestra normativa.

Método de elección forzosa. En el desarrollo de pruebas de personalidad, método de redacción de los reactivos en los que se equiparan las alternativas en cuanto a deseabilidad social.

Método de escalamiento absoluto. Procedimiento para obtener una medida de la dificultad absoluta con base en los resultados de grupos de individuos de diversas edades.

Método de escalamiento racional. Método de construcción de escalas en el que todos los reactivos se correlacionan en sentido positivo unos con otros y también con la calificación total de la escala; asimismo se le conoce como enfoque de consistencia interna.

Método de intervalos de aparición balanceada. Método para construir escalas a nivel intervalo a partir de afirmaciones sobre actitud.

Moda. Puntuación de ocurrencia más frecuente.

Modelo integrativo. Modelo de evaluación de carrera en el que se considera de manera simultánea la información sobre las áreas de intereses, capacidad y personalidad.

Modelo RIASEC. Teoría de tipos persona-ambiente que propone seis temas: Realista, Investigador, Artístico, Social, Emprendedor y Convencional (RIASEC).

Muestra conductual. En pruebas, el concepto de que una prueba sólo es una muestra de conductas que permite que el evaluador haga inferencias acerca de un dominio más amplio de conductas relacionadas.

Muestra de estandarización. Grupo de gran tamaño de individuos representativos de la población a la que se dirige la prueba.

Muestra de trabajo. Procedimiento de evaluación que utiliza una réplica en miniatura del trabajo que han solicitado los individuos examinados.

Muestreo aleatorio estratificado. Estrategia de selección en la que los individuos se eligen de manera aleatoria, con restricciones de que la muestra se equipare a la población en cuanto a variables antecedentes relacionadas como raza, sexo, ocupación, etc.

Muestreo aleatorio. Estrategia de selección en la que existe igual probabilidad de selección para todos los individuos.

Multimedia. Capacidad colectiva de las computadoras modernas para utilizar imágenes fijas, segmentos de vídeo en vivo, música, cuadros, gráficas, animación y otros enfoques en un formato interactivo.

Neuropsicología. Estudio de la relación entre función cerebral y conducta.

Nivel basal. Para las pruebas en las que los reactivos de una subprueba se distribuyen del más fácil al más difícil, es el nivel bajo al cual es casi seguro que el individuo evaluado responderá correctamente a todas las preguntas.

Nivel superior. Para las pruebas en las que los reactivos de una subprueba se distribuyen del más fácil al más difícil, el nivel por en-

cima del cual es casi seguro que el individuo evaluado fallará en todas las siguientes preguntas.

No culpable en razón de insania (NCRI). Un veredicto que se admite en algunos estados de EUA en el que se encuentra no culpable al acusado debido a haber cometido el acto criminal como resultado de enfermedad o defecto mental.

Norma de grado. Un tipo de estandarización que representa el nivel de desempeño en una prueba para cada grado escolar por separado en una muestra normativa.

Norma por edad. Tipo de estandarización que representa el nivel de desempeño en una prueba para cada grupo independiente por edad, dentro de la muestra normativa.

Normas de subgrupo. Normas obtenidas de un subgrupo identificado, en contraposición a una muestra nacional diversificada.

Normas locales. Normas obtenidas de una muestra local representativa, en contraposición a una muestra nacional.

Normas. Resumen de resultados de una prueba para un grupo amplio y representativo de personas.

Ojiva normal. Distribución normal graficada en forma acumulativa.

Organización perceptual. Segundo factor en el WISC-III que incluye Ordenamiento de Dibujos, Figuras Incompletas, Diseños con Cubos y Ensamble de Objetos.

Paralingüística. Aspectos no verbales del habla, como tono de voz y tasa de expresión verbal.

Pares craneales. Doce pares de vías neuronales que ayudan a gobernar las funciones sensoriales y motoras básicas, como visión, olfato, movimientos faciales, gusto y audición.

Patrón conductual tipo A con propensión a la cardiopatía. Patrón de conducta consistente en inseguridad acerca del estatus, agresividad excesiva, hostilidad libre flotante y un sentido de urgencia en cuanto al tiempo (enfermedad de la prisa).

Patrones de reforzadores ocupacionales. Valoración de los puestos en términos de los reforzadores percibidos por el trabajador, que están presentes o ausentes.

Pensamiento divergente. Tipo de pensamiento que va en diferentes direcciones.

Percentil. Porcentaje de personas en la muestra de estandarización que tuvieron calificaciones por debajo de una puntuación natural específica; los percentiles varían de 0 a 100.

Personalidad. Constructo no explícito que se usa para explicar la consistencia conductual dentro de las personas y la inconsistencia conductual entre individuos.

Polígono de frecuencias. Método para resumir los datos o calificaciones de una prueba en forma de gráfica; similar a un histograma, excepto que la frecuencia de los intervalos de clase se representa con puntos en vez de columnas.

Polígrafo. Instrumento que registra respuestas fisiológicas en proceso, incluyendo cambios en respiración, frecuencia respiratoria, presión sanguínea y transpiración; conocido de manera inadecuada como "detector de mentiras".

Preocupaciones externas a la validez. Efectos colaterales y consecuencias no intencionales de las pruebas.

Problema de criterio. El difícil problema de conceptuar y medir constructos relacionados con el desempeño laboral que con frecuencia son complejos, confusos y multidimensionales.

Procedimiento conductual. Procedimiento para evaluar los antecedentes y consecuencias de la conducta; incluye listas de verificación, escalas de calificación, entrevistas y observaciones estructuradas.

Procedimiento de estandarización. En pruebas, el intento, a través de instrucciones cuidadosamente estrictas, de asegurar que los

procedimientos para aplicación de una prueba son uniformes de un examinador y ambiente a otro.

Procedimiento de manejo de contingencias. Enfoque de terapia conductual en el que el terapeuta identifica y altera las consecuencias de las conductas indeseables.

Procesamiento simultáneo. Forma de procesamiento de información caracterizada por la ejecución simultánea de varias operaciones mentales diferentes.

Procesamiento sucesivo. Forma de procesamiento de información en el que debe seguirse una secuencia apropiada de operaciones mentales.

Producción divergente. Creación de numerosas respuestas apropiadas para una sola situación estímulo.

Propensión de respuesta de deseabilidad social. Tendencia de las personas examinadas a reaccionar ante la deseabilidad (o indeseabilidad) percibida de un reactivo de prueba en lugar de responder de modo preciso a su contenido.

Prosencéfalo. Porción grande, externa del cerebro que incluye a la corteza cerebral y estructuras subyacentes como el cuerpo calloso, ganglios basales, lóbulo límbico, tálamo e hipotálamo.

Prueba. Procedimiento estandarizado para tomar una muestra de conducta y describirla con categorías o calificaciones. Además, la mayor parte de las pruebas tienen normas o estándares con base en los cuales pueden utilizarse los resultados para predecir otras conductas más importantes.

Prueba adaptativa computarizada. Familia de procedimientos que permiten la evaluación precisa y eficaz de la capacidad; se continúa con la prueba individualizada hasta que se alcanza un nivel predeterminado de precisión en la medición.

Prueba culturalmente justa. Prueba diseñada para reducir al mínimo las influencias del aprendizaje cultural y del clima social y, por tanto, producir una separación más definida entre la capacidad natural y el aprendizaje específico.

Prueba de aptitud. Prueba que mide uno o más segmentos definidos de modo claro y relativamente homogéneos de capacidad.

Prueba de creatividad. Prueba que evalúa la capacidad para producir nuevas ideas, discernimientos o creaciones artísticas aceptadas como valiosas a nivel social, estético o científico.

Prueba de evitación conductual. Procedimiento conductual en el que el terapeuta mide cuánto tiempo el individuo puede tolerar un estímulo productor de ansiedad.

Prueba de integridad abierta. Prueba de empleo que busca evaluar las actitudes hacia el robo; estos instrumentos pueden contener también una sección que trata sobre la admisión abierta del mismo.

Prueba de integridad. Instrumento diseñado para detectar a empleados potenciales en cuanto a propensión al robo y otras cualidades indeseables; las pruebas explícitas de integridad contienen preguntas sobre actitudes en cuanto al robo y reactivos que tratan de la admisión del robo y de otras actividades ilegales.

Prueba de inteligencia. Aunque existen excepciones, en general una prueba de este tipo produce una calificación general resumida que se basa en los resultados de una muestra heterogénea de reactivos (p. ej., habilidades verbales, razonamiento, pensamiento espacial).

Prueba de personalidad. Prueba que mide los rasgos cualidades o conductas que determinan la individualidad de una persona; esta información ayuda a predecir la conducta futura.

Prueba de poder. Prueba que concede el suficiente tiempo a la persona examinada para hacer el intento de resolver todos los reactivos; no obstante, la prueba tiene la

suficiente dificultad como para que ninguno de los individuos que la responda pueda obtener una calificación perfecta.

Prueba de rendimiento. Prueba que mide el grado de aprendizaje, éxito o logro en una materia.

Prueba de velocidad. Prueba cronometrada que contiene reactivos de nivel de dificultad uniforme y simple en términos generales; el límite de tiempo es suficientemente estricto como para que pocos individuos terminen la prueba cronometrada.

Prueba directiva. Subprueba inicial utilizada para determinar el nivel de entrada para todas las subpruebas restantes; utilizada con pruebas individuales de inteligencia como la SB:CE.

Prueba ipsativa. Prueba en la que el promedio de las subescalas es siempre el mismo para todos los individuos examinados; así, para uno determinado, las calificaciones altas en las subescalas deben equilibrarse con aquellas bajas en otras subescalas.

Prueba proyectiva. Prueba en la que la persona evaluada se enfrenta con estímulos vagos y ambiguos, y responde con sus propias construcciones.

Prueba referida a criterio. Prueba en la que el objetivo consiste en determinar en qué sitio se coloca la persona evaluada con respecto a objetivos educativos definidos de manera muy estrecha; no se hace ninguna comparación con el desempeño de otros individuos evaluados.

Prueba referida a normas. Prueba en la que el desempeño de cada individuo examinado se interpreta con referencia a una muestra relacionada de estandarización.

Pruebas de rendimiento individual. Pruebas aplicadas de modo individual para estimar los niveles de rendimiento; éstas son esenciales en la evaluación de posibles discapacidades para el aprendizaje.

Pruebas grupales. Medidas principalmente de lápiz y papel adecuadas para examinar a grandes grupos de personas al mismo tiempo.

Pruebas grupales de rendimiento. También llamadas pruebas de rendimiento educativo, estos instrumentos se aplican por lo común a docenas o cientos de estudiantes al mismo tiempo para estimar los niveles de rendimiento en una o más áreas académicas bien definidas.

Pruebas individuales. Instrumentos que por su diseño y propósito deben aplicarse a una persona.

Pruebas neuropsicológicas. Pruebas y procedimientos con sensibilidad probada a los efectos del daño cerebral.

Psicología del consumo. Rama de la psicología industrial y organizacional que trata del desarrollo, promoción y mercadeo de productos y servicios.

Psicometrista. Especialista en psicología o educación que desarrolla y evalúa pruebas psicológicas.

Puntuación estándar normalizada. Puntuación obtenida por una transformación que traduce una curva sesgada en una distribución normal.

Puntuación estándar. Puntuación transformada en la que la calificación original se expresa como la distancia con respecto a la media en unidades de desviación estándar.

Puntuación natural. Nivel más básico de la información proporcionada por una prueba psicológica; por ejemplo, número de preguntas respondidas de manera correcta.

Puntuación T. Puntuación transformada con una media de 50 g desviación estándar de 10.

Puntuación verdadera. La puntuación hipotética real de un individuo examinado en una prueba; la puntuación verdadera puede estimarse de modo probabilístico, pero nunca se conoce directamente.

Rapport. En pruebas, una atmósfera cómoda y cálida que sirve para motivar a los individuos examinados y producir cooperación.

Rasgo. Cualquier manera relativamente duradera en la que un individuo difiere de otro.

Rasgos fuente. Fuentes estables y constantes de la conducta que son menos visibles que los rasgos superficiales, pero más importantes para explicar la conducta.

Rasgos superficiales. En la teoría de Cattell, los aspectos más obvios de la personalidad que surgen por lo común en las primeras etapas del análisis factorial cuando los reactivos individuales de una prueba se correlacionan entre sí.

Realidad virtual. Utilización de imágenes elaboradas por computadora que se proyectan en gafas que rodean la cabeza para representar un ambiente tridimensional, cambiante, en movimiento.

Reducción de la validez. Descubrimiento común en la investigación de validación cruzada de que una prueba predice de manera menos precisa el criterio relacionado con la nueva muestra de personas evaluadas que con la del piloteo original.

Regla del Código Penal Modelo. Estándar de demencia legal – "Una persona no es responsable de conducta criminal si al momento de dicha conducta, como resultado de enfermedad o defecto mental, carece de la capacidad sustancial ya sea para apreciar la criminalidad [incorrección] de su conducta o para conformar su conducta a los requisitos de la ley".

Regla Durham. La disposición legal para la defensa de insania si el acto criminal fue "producto" de enfermedad o defecto mental; se descartó en 1972 y se le reemplazó con el Código Penal Modelo.

Religión como Búsqueda. La perspectiva de que la complejidad, duda y vacilación son aspectos de la expresión religiosa madura.

Restricción del rango. Fenómeno en el que se restringe el rango de una variable, lo cual produce que las correlaciones con otras variables sean artificialmente bajas.

Retraso mental. Funcionamiento intelectual general que se encuentra significativamente por debajo del promedio y que resulta en o se relaciona con la discapacidad concurrente en conducta adaptativa y se manifiesta durante el periodo del desarrollo.

Rotación a estructura simple. En análisis factorial, método para rotar la matriz factorial que busca simplificar las cargas factoriales de modo que cada prueba tenga cargas significativas en el menor número de factores posible.

Rotación a múltiple positivo. En análisis factorial, método para rotar la matriz factorial que busca eliminar tantas cargas factoriales negativas como sea posible.

Sesgo de una prueba. En su uso popular, una prueba está sesgada si discrimina de manera injusta en contra de minorías raciales y étnicas, mujeres y personas en situación de pobreza; a nivel técnico, el sesgo de una prueba se refiere a la validez diferencial para subgrupos definibles y relacionados de personas.

Sesgo del calificador. Tendencia a que las calificaciones del supervisor sean imprecisas debido a indulgencia, severidad y otros errores de medición.

Sesgo en la validez de constructo. Tipo de sesgo que ocurre cuando se demuestra que una prueba mide diferentes rasgos hipotéticos (constructos psicológicos) para un grupo que para otro o mide el mismo rasgo, pero con distintos grados de precisión.

Sesgo en la validez de contenido. Tipo de sesgo que se demuestra cuando un reactivo o subescala es relativamente más difícil para los miembros de un grupo que para los de otro, después de mantener constante el nivel general de capacidad de ambos grupos.

Sesgo en la validez predictiva. Tipo de sesgo demostrado cuando la inferencia obtenida de la calificación de una prueba no se realiza con el menor error aleatorio razonable o existe un error constante en la inferencia o predicción en función de la membresía a un grupo en particular.

Sesgo. La simetría o asimetría de una distribución de frecuencias; el sesgo positivo indica que las calificaciones se encuentran agrupadas en el extremo bajo, el sesgo negativo indica que éstas se encuentran agrupadas en el extremo alto.

Síndrome alcohólico fetal. Conjunto de anormalidades físicas y conductuales que incluyen retraso mental, producidas por el hecho de que la madre ingiera alcohol durante el embarazo.

Superyó. En la teoría psicoanalítica, aquella parte de la personalidad que es aproximadamente sinónimo de conciencia y que incluye las normas sociales de lo que es correcto e incorrecto trasmitidas por los padres.

Tabla de especificaciones. En el desarrollo de pruebas, un cuadro que lista el número exacto de reactivos en áreas de contenido relacionadas; una tabla de este tipo también especifica el número preciso de reactivos que deben representar diferentes procesos cognitivos.

Tálamo. Estructura clave que proporciona entrada sensorial e información acerca de movimiento en progreso a la corteza cerebral; el tálamo es la principal estación de relevo en el cerebro.

Talentos especiales. La designación de una persona como talentosa significa que por lo común tiene una capacidad extraordinaria en alguna área.

Tasa base. En la teoría de decisión, la proporción de solicitantes exitosos que pueden seleccionarse utilizando los métodos habituales, sin el beneficio de la nueva prueba.

Técnica de bandeja de entrada. Prueba de una muestra realista de trabajo que simula el ambiente laboral de un administrador.

Técnica Q. Técnica para el estudio de cambios en el autoconcepto y otras variables mediante la clasificación de afirmaciones en una distribución casi normal para categorías asignadas.

Teoría clásica de medición. Teoría dominante en las pruebas psicológicas; supone que una calificación observada consta de una calificación verdadera más un error de medición.

Teoría de decisión. Enfoque de medición psicológica que considera los costos y beneficios de las decisiones basadas en la prueba; por ejemplo, en la selección de personal.

Teoría de la capacidad para la generalización. Modelo de confiabilidad del muestreo de un dominio que reconoce varias alternativas para la generalización de resultados de una prueba.

Terapia cognitiva conductual. Enfoque de cambio conductual que enfatiza la modificación de la estructura de creencias del cliente.

Terapia conductual. La aplicación de los métodos y descubrimientos de la psicología experimental para la modificación de la conducta desadaptada; también se le llama modificación conductual.

Teratógeno. Sustancia que atraviesa la barrera placentaria y produce deformidades físicas en el feto.

Testigo experto. En juicios legales, un testigo que el juez considera calificado para testificar acerca de un tema determinado.

Trastorno por déficit de atención con hiperactividad. Síndrome conductual caracterizado por intranquilidad, distracción, impulsividad, déficit de atención, habilidades sociales deficientes e incapacidad para considerar las consecuencias.

Validación cruzada. En el caso de pruebas predictivas, la práctica de utilizar la ecuación de regresión original en una nueva muestra para determinar si la prueba predice el criterio tan bien como lo hizo en la muestra original.

Validación ilusoria. En pruebas proyectivas, el descubrimiento de que los individuos ignoran las instancias de desconfirmación y se adhieren a sus estereotipos preexistentes.

Validez concurrente. Tipo de validez relacionada con el criterio en la que se obtienen las medidas de criterio aproximadamente al mismo tiempo que las calificaciones de una prueba.

Validez convergente. Tipo de validez que se demuestra cuando una prueba se correlaciona en gran medida con otras variables o pruebas con las que comparte una superposición de constructos.

Validez de constructo. Tipo de validez que se refiere a la idoneidad de las inferencias basadas en la prueba en cuanto al constructo subyacente que ésta se propone medir.

Validez de contenido. Tipo de validez que se determina por el grado en que las preguntas, tareas o reactivos de una prueba son representativos del universo de conducta del que ésta tiene el propósito de ser una muestra.

Validez de discriminación. Tipo de validez que se demuestra cuando una prueba no se correlaciona con variables o pruebas con las cuales debería diferir.

Validez de Facie. En el caso de pruebas, apariencia de validez para los usuarios de la prueba, examinadores y, en particular, para los individuos evaluados; no es una forma técnica de validez, pero es importante para la aceptabilidad social de una prueba.

Validez predictiva. Tipo de validez relacionada con el criterio en el que las medidas de criterio se obtienen en el futuro, en general meses o años después de que se obtuvieron las calificaciones; por ejemplo, cuando se pronostican las calificaciones en la universidad a partir del examen de ingreso.

Validez relacionada con el criterio. Tipo de validez que se demuestra cuando se pone en evidencia que una prueba es eficaz en la estimación del desempeño de una persona evaluada en alguna medida de resultado.

Validez. Una prueba es válida al grado en que las inferencias realizadas con base en ella sean apropiadas, significativas y útiles.

Valor. Según Rokeach y otros investigadores, una creencia compartida y duradera acerca de los modos ideales de conducta o estados meta de la existencia.

Valoración para custodia. En casos de divorcio la valoración psicológica de un niño (o niños) y de ambos padres para ofrecer una opinión ante el tribunal sobre los intereses primordiales de éste (o éstos) en relación con los arreglos de custodia.

Valores laborales. Necesidades, motivos y valores que influyen la elección vocacional, la satisfacción con el puesto y el desarrollo de la carrera.

Varianza. Índice estadístico que refleja el grado de dispersión en un grupo de calificaciones.

Velocidad de procesamiento. Cuarto factor del WISC-III que incluye Claves y Símbolos, y Dígitos.

Ventrículos. Cavidades llenas de líquido dentro del cerebro.

Yo. En la teoría psicoanalítica, sí mismo consciente que media entre el ello y la realidad. ■

Referencias

Aamodt, M. G., Keller, R., Crawford, K., & Kimbrough, W. (1981). A critical incident job analysis of the university housing resident assistant position. *Psychological Reports, 49*, 983-986.

Abel, E. L. (1995). An update on incidence of FAS:FAS is not an equal opportunity birth defect. *Neurobehavioral Toxicology, 17*, 427-443.

Abell, S. C., Briesen, P., & Watz, L. (1996). Intellectual evaluations of children using human figure drawings: An empirical investigation of two methods. *Journal of Clinical Psychology, 52*, 67-74.

Achenbach, T. M. (1991). *Manual for the Teacher's Report Form and 1991 Profile.* Burlington, VT: University of Vermont, Department of Psychiatry.

Achenbach, T. M. (1992). *Manual for the Child Behavior Checklist/2-3 and 1992 Profile.* Burlington, VT: University of Vermont, Department of Psychiatry.

Ackerman, M. J., & Schoendorf, K. (1992). *Ackerman-Schoendorf Scales for Parent Evaluation of Custody (ASPECT).* Los Angeles: Western Psychological Services.

Ackerman, P. L. (1992). Review of Stanford Early School Achievement Test. *Eleventh Mental Measurements Yearbook.* Lincoln, NE: University of Nebraska Press.

Adams, G. A., Elacqua, T., & Colarelli, S. (1994). The employment interview as a sociometric selection technique. *Journal of Group Psychotherapy, Psychodrama, and Sociometry.* Fall, 99-113.

Adams, K. M., and Heaton, R. K. (1985). Automated interpretation of neuropsychological test data. *Journal of Consulting and Clinical Psychology, 53*, 790-802.

Ahrens, J., Evans, R., & Barnett, R. (1990). Factors related to dropping out of school in an incarcerated population. *Educational and Psychological Measurement, 50,* 611-617.

Aiken, L. R. (1989). *Assessment of personality.* Boston: Allyn & Bacon.

Aita, J. A., Armitage, S. G., Reitan, R. M., & Rabinowitz, A. (1947). The use of certain psychological tests in the evaluation of brain injury. *Journal of General Psychology, 37,* 25-44.

Albert, M., Sparks, R., von Stockert, T., & Sax, D. (1972). A case study of auditory agnosia: linguistic and non-linguistic processing. *Cortex, 8,* 427-455.

Albert, S., Fox, H. M., & Kahn, M. W. (1980). Faking psychosis on the Rorschach: Can expert judges detect malingering? *Journal of Personality Assessment, 44,* 115-119.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory.* Monterey, CA: Brooks/Cole.

Allport, G. W. (1937). *Personality: A psychological interpretation.* New York: Holt, Rinehart & Winston.

Allport, G., & Ross, J. (1967). Personal religious orientation and prejudice. *Journal of Personality and Social Psychology, 5,* 432-443.

Allport, G., Vernon, P. E., & Lindzey, G. (1960). *Study of Values* (3rd ed.): Manual. Chicago: Riverside Publishing Company.

Allport, G. W. (1950). *The individual and his religion.* New York: Macmillan.

Allport, G. W., & Odbert, H. (1936). Trait names, a psycholexical study. *Psychological Monographs, 47* (Whole No. 211).

Allport, G. W., & Vernon, P. E. (1931). *Study of values: A scale for measuring the dominant interests in personality.* Boston: Houghton Mifflin.

Altepeter, T. S. (1989). The PPVT-R as measure of psycholinguistic functioning: A caution. *Journal of Clinical Psychology, 45,* 935-941.

Altepeter, T. S., & Johnson, K. A. (1989). Use of the PPVT-R for intellectual screening with adults: A caution. *Journal of Psychoeducational Assessment, 7,* 39-45.

Amabile, T. M. (1983). *The social psychology of creativity.* New York: Springer-Verlag.

Amelang, M., & Borkenau, P. (1986). The trait concept: Current theoretical considerations, empirical facts, and implications for personality inventory construction. In A. Angleitner & J. S. Wiggins (eds.), *Personality assessment via questionnaires: Current issues in theory and measurement.* Berlin: Springer-Verlag.

American Association for Counseling and Development. (1988). *Ethical standards.* Washington, DC: Author.

American Association on Mental Retardation. (1992). *Mental retardation: Definition, classification, and systems of supports.* Washington, DC: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (2nd ed.). Washington, DC: American Psychological Association.

American Federation of Teachers, National Council on Measurement in Education, and National Education Association. (1990). *Standards for teacher competence in educational assessment of students.* Washington, DC: Author.

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

American Psychological Association. (1953). *Ethical standards of psychologists.* Washington, DC: Author.

American Psychological Association. (1986). *Guidelines for computer-based tests and interpretations.* Washington, DC: Author.

American Psychological Association. (1988). In the Supreme Court of the United States: *Clara Watson v. Fort Worth Bank & Trust. American Psychologist, 43,* 1019-1028.

American Psychological Association. (1992a). Ethical principles of psychologists and code of conduct. *American Psychologists, 47,* 1597-1611.

American Psychological Association. (1992b). *Psychological testing of language minority and culturally different children.* Washington, DC: Author.

American Psychological Association. (1993). Guidelines for providers of psychological services to ethnic, linguistic, and culturally diverse populations. *American Psychologist, 48,* 45-48.

American Psychological Association. (1994a). Guidelines for child custody evaluations in divorce proceedings. *American Psychologist, 49,* 677-680.

American Psychological Association. (1994b). Report of the ethics committee, 1993. *American Psychologist, 49,* 659-666.

American Speech-Language-Hearing Association. (1991). *Code of ethics of the American Speech-Language Hearing Association.* Rockville, MD: Author.

Anastasi, A. (1975). Review of the Goodenough-Harris Drawing Test. *The Seventh Mental Measurements Yearbook.* Lincoln, NE: University of Nebraska Press.

Anastasi, A. (1981). Coaching, test sophistication, and developed abilities. *American Psychologist, 36,* 1086-1093.

Anastasi, A. (1985). *Psychological testing* (6th ed.). New York: Macmillan.

Anastasi, A. (1986). Emerging concepts of test validation. *Annual Review of Psychology, 37*, 1-15.

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York, NY: Macmillan.

Andersson, H. W. (1996). The Fagan Test of Infant Intelligence: Predictive validity in a random sample. *Psychological Reports, 78*, 1015-1026.

Andrew, D. M., Peterson, D. G., Longstaff, H. P. (1979). *Minnesota Clerical Test Manual*. San Antonio, TX: The Psychological Corporation.

Ansorge, C. J. (1985). Review of the Cognitive Abilities Test. *Ninth Mental Measurements Yearbook*. Lincoln, NE: University of Nebraska Press.

Anthony, J. C., LeResche, L., Niaz, U., Von Korff, M., & Folstein, M. (1982). Limits of the Mini-Mental State as a screening test for dementia and delirium among hospital patients. *Psychological Medicine, 12*, 397-408.

Antoni, M. (1993). The combined use of the MCMI and MMPI. In R. J. Craig (ed.), *The Millon Clinical Multiaxial Inventory: A clinical research information synthesis*. Hillsdale, NJ: Erlbaum.

Argentero, P. (1989). Second-order factor structure of Cattell's 16 Personality Factor Questionnaire. *Perceptual and Motor Skills, 68*, 1043-1047.

Argyle, M., & Cook, M. (1976). *Gaze and mutual gaze*. Cambridge: Cambridge University Press.

Arthur, G. (1950). *The Arthur adaptation of The Leiter International Performance Scale*. Chicago: Stoelting Co.

Arvey, R. D., & Begalla, M. E. (1975). Analyzing the homemaker job using the Position Analysis Questionnaire (PAQ). *Journal of Applied Psychology, 60*, 513-517.

Arvey, R. D., & Campion, J. E. (1982). The employment interview: A summary and review of recent research. *Personnel Psychology, 35*, 281-332.

Arvey, R. D., & Faley, R. H. (1988). *Fairness in selecting employees*. Reading, MA: Addison-Wesley.

Arvey, R. D., & Murphy, K. R. (1998). Performance evaluation in work settings. *Annual Review of Psychology, 49*, 141-168.

Arvey, R. D., Landon, T., Nutting, S., & Maxwell, S. (1992). Development of physical ability tests for police officers: A construct validation approach. *Journal of Applied Psychology, 77*, 996-1009.

Asher, J. J., & Sciarrino, J. A. (1974). Realistic work samples: A review. *Personnel Psychology, 27*, 519-533.

Auer, S., & Reisberg, B. (1996). Reliability of the modified Ordinal Scales of Psychological Development: A cognitive assessment battery for severe dementia. *International Psychogeriatrics, 8*, 225-231.

Austin, J. T., & Villanova, P. (1992). The criterion problem: 1917-1992. *Journal of Applied Psychology, 77*, 836-874.

Axelrod, . N., Greve, K., & Goldman, R. (1994). Comparison of four Wisconsin Card Sorting Test Scoring guides with novice raters. *Assessment, 1*, 115-121.

Bach, P. J., Harowski, K., Kirby, K., Peterson, P., & Schulein, M. (1981). The interrater reliability of the Luria-Nebraska Neuropsychological Battery. *Clinical Neuropsychology, 3*, 19-21.

Bagby, R. M., Nicholson, R., Rogers, R., and Nussbaum, D. (1992). Domains of competency to stand trial: A factor analytic study. *Law and Human Behavior, 16*, 491-507.

Bagby, R. M., Rogers, R., Buis, T., & Kalemba, V. (1994). Malingered and defensive response styles on the MMPI-2: An examination of validity scales. *Assessment, 1*, 31-38.

Ballard, J., & Zettel, J. (1977). Public Law 94-142 and Sec. 504: What they say about rights and protections. *Exceptional Children, 44*, 177-185.

Ball-Rokeach, S., Rokeach, M., & Grube, J. (1984). *The great American values test*. New York: Free Press.

Baltes, P. B., Reese, H., & Nesselroade, J. (1977). *Lifespan developmental psychology: Introduction to research methods*. Belmont, CA: Wadsworth.

Balthazar, E. E. (1976). *Balthazar Scales of Adaptive Behavior*. Palo Alto, CA: Consulting Psychologists Press.

Bandura, A. (1965). Vicarious processes: A case of notrial learning. In L. Berkowitz (ed.), *Advances in experimental social psychology* (vol. 2). New York: Academic Press.

Bandura, A. (1971). *Social Learning theory*. Morristown, NJ: General Learning Press.

Bandura, A. (1977). *Social learning*. Englewood Cliffs, NJ: Prentice-Hall.

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist, 37*, 122-147.

Bandura, A., & Walters, R. H. (1963). *Social learning and personality development*. New York: Holt, Rinehart & Winston.

Bandura, A., Taylor, C. B., Ewart, C. K., Miller, N. M., & BeBusk, R. F. (1985). Exercise testing to enhance wives' confidence in their husbands' cardiac capability soon after clinically uncomplicated acute myocardial infarction. *American Journal of Cardiology, 55*, 635-638.

Banerji, M. (1992). An integrated study of the predictive properties of the Gesell School Readiness Test. *Journal of Psychoeducational Assessment, 10*, 240-256.

Barkley, R. A. (1981). *Hyperactive children: A handbook for diagnosis and treatment*. New York: Guilford Press.

Barkley, R. A., & Edelbrock, C. S. (1987). Assessing situational variation in children's behavior problems: The home and school situations questionnaires. In R. Prinz (ed.), *Advances in behavioral assessment of children and families* (vol. 3). Greenwich, CT: JAI Press.

Barron, F. (1953). An ego-strength scale wich predicts response to psychotherapy. *Journal of Consulting Psychology, 17*, 327-333.

Batson, C. D., Schoenrade, P., & Ventis, W. (1993). *Religion and the individual: A Social-psychological perspective*. New York: Oxford University Press.

Bausell, R. B. (1986). *A practical guide to conducting empirical research*. New York: Harper & Row.

Bayless, J. D., Varney, N. R., & Roberts, R. J. (1989). Tinker Toy Test performance and vocational outcome in patients with closed-head injuries. *Journal of Clinical and Experimental Neuropsychology, 11*, 913-917.

Bayley, N. (1969). *Bayley Scales of Infant Development*. San Antonio, TX: The Psychological Corporation.

Bayley, N. (1993). *Bayley Scales of Infant Development-II*. San Antonio, TX: The Psychological Corporation.

Bear, D. M. (1986). Behavioural changes in temporal lobe epilepsy. In M. R. Trimble & T. G. Bolwig (eds.), *Aspects of epilepsy and psychiatry*. New York: Wiley.

Beck, A. T. (1976). *Cognitive therapy and the emotional disorders*. New York: New American Library.

Beck, A. T. (1983). Negative cognitions. In E. Levitt, B. Lubin, & J. Brooks (eds.), *Depression: Concepts, controversies, and some new facts* (2nd ed.). Hillsdale, NJ: Erlbaum.

Beck, A. T. (1987). Cognitive models of depression. *Journal of Cognitive Psychotherapy, An International Quarterly, 1*, 5-37.

Beck, A. T., & Steer, R. A. (1987). *Manual for the revised Beck Depression Inventory*. San Antonio, TX: Psychological Corporation.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for the Beck Depression Inventory-II*. San Antonio, TX: Psychological Corporation.

Beck, A. T., Steer, R. A., & Garbin, M. G. (1988). Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clinical Psychology Review, 8*, 77-100.

Beck, A. T., Ward, C. H., Mendelsohn, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4*, 561-571.

Becker, T. E., & Colquitt, A. L. (1992). Potential versus actual faking of a biodata form: An analysis along several dimensions of item type. *Personnel Psychology, 45*, 389-406.

Belcher, M. J. (1992). Review of the Wonderlic Personnel Test. *The Eleventh Mental Measurements Yearbook*. Lincoln, NE: University of Nebraska Press.

Bellak, L. (1992). *The Thematic Apperception Test, the Children's Apperception Test, and the Senior Apperception Technique in clinical use* (5th ed.). Orlando, FL: Grune & Stratton.

Bellak, L., & Bellak, S. S. (1991). *Children's Apperception Test Manual (CAT)* (8th rev. ed.). Larchmont, NY: C. P. S.

Bellak, L., & Bellak, S. S. (1994). *Children's Apperception Test Human Figures (CAT-H)* (11th ed.). Larchmont, NY: C. P. S.

Bem, D. J., & Allen, A. (1974). On predicting some of the people some of the time: The search for cross-situational consistencies in behavior. *Psychological Review, 81*, 506-520.

Bem, D., & Funder, D. (1978). Predicting more of the people more of the time: Assessing the personality of situations. *Psychological Review, 85*, 485-501.

Bender, L. (1938). *A visual motor gestalt test and its clinical use*. New York: American Orthopsychiatric Association.

Bennett, G. K., Seashore, H. G., & Wesman, A. G. (1974). *Fifth edition manual for the Differential Aptitude Tests, Forms S and T*. San Antonio, TX: Psychological Corporation.

Bennett, G. K., Seashore, H. G., & Wesman, A. G. (1982). *Differential Aptitude Tests: Administrator's handbook*. San Antonio, TX: Psychological Corporation.

Bennett, G. K., Seashore, H. G., & Wesman, A. G. (1984). *Differential Aptitude Tests: Technical Supplement*. San Antonio, TX: Psychological Corporation.

Bennett, T. (1988). Use of the Halstead-Reitan Neuropsychological Test Battery in the assessment of head injury. *Cognitive Rehabilitation, 6*, 18-25.

Ben-Porath, Y. S. (1994). The ethical dilemma of coached malingering research. *Psychological Assessment, 6*, 16-17.

Ben-Porath, Y. S., & Butcher, J. N. (1989). Psychometric stability of rewritten MMPI items. *Journal of Personality Assessment, 53*, 645-653.

Benson, D. F. (1988). Disorders of visual gnosis. In J. W. Brown (eds.), *The neuropsychology of visual perception*. Hillsdale, NJ: Erlbaum.

Benson, D. F. (1994). *The neurology of thinking*. New York: Oxford University Press.

Benson, P. G. (1985). Minnesota Importance Questionnaire. In D. J. Keyser & R. C. Sweetland (eds.), *Test critiques* (vol. 2). Kansas City, MO: Test Corporation of America.

Benson, P., Donahue, M., & Erickson, J. (1993). The Faith Maturity Scale: Conceptualization, Measurement, and Empirical Validation. In M. L. Lynn & D. O. Moberg (eds.), *Research in the social scientific study of religion* (vol. 5). Greenwich, CN: JAI Press.

Benton, A., Hamsher, K., Rey, G., & Sivan, A. (1994). *Multilingual Aphasia Examination* (3rd ed.). Iowa City, IA: AJA Associates.

Benton, A., Sivan, A., Hamsher, K., Varney, N., & Spreen, O. (1994). *Contributions to neuropsychological assessment* (2nd ed). New York: Oxford University Press.

Bentz, V. J. (1985). Research findings from personality assessment of executives. In H. J. Bernardin & D. A. Bownas (eds.), *Personality assessment in organizations*. New York: Praeger.

Berg. E. A. (1948). A simple objective test for measuring flexibility in thinking. *Journal of General Psychology, 39*, 15-22.

Berger, S. G., Chibnall, J., & Gfeller, J. (1994). The Category Test: A comparison of computerized and standard versions. *Assessment, 3*, 255-258.

Berk, L. E. (1989). *Child development*. Boston: Allyn & Bacon.

Berk, R. A. (ed.). (1984). *A guide to criterion-referenced test construction*. Baltimore: Johns Hopkins University Press.

Bernreuter, R. G. (1931). *The personality inventory*. Stanford, CA: Stanford University Press.

Bersoff, D. N. (1984). Social and legal influences on test development and usage. In B. S. Plake (ed.), *Social and technical issues in testing: Implications for test construction and usage*. Hillsdale, NJ: Erlbaum.

Bersoff, D. N. (1988). Should subjective employment devices be scrutinized? Its elementary, my dear Ms. Watson. *American Psychologist, 43*, 1016-1018.

Bigler, E. D. (ed.). (1990). *Traumatic brain injury: Mechanisms of damage, assessment, intervention, and outcome*. Austin, TX: PRO-ED.

Binet, A., & Simon, T. (1905). Methodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *Annee Psychologique, 11*, 191-244.

Bishop, E. G., Hailey, B. J., & O'Rourke, D. F. (1989). Reliability of the Jenkins Activity Survey—Form T: Temporal stability and internal consistency. *Journal of Personality Assessment, 53*, 60-65.

Bishop, J., Knights, R. M., Stoddart, C. (1990). Rey Auditory-Verbal Learning Test: Performance of English and French children aged 5 to 16. *The Clinical Neuropsychologist, 4*, 133-140.

Blake, R. J., Potter, E., III, Sliwak, R. (1993). Validation of the structural scales of the CPI for predicting the performance of junior officers in the U.S. Coast Guard. *Journal of Business Psychology, 7*, 431-448.

Blakley, B. R., Quinones, M., Crawford, M., & Jago, I. (1994). The validity of isometric strength tests. *Personnel Psychology, 47*, 247-274.

Blakely, T., Jr., Crinella, F., Fisher, T., Champaigne, L., & Beck, F. (1994). Neuropsychological correlates of learning disabilities: Identification by the Tryon clustering method. *Journal of Developmental and Physical Disabilities, 6*, 1-22.

Blau, T. (1984). *The psychologist as expert witness*. New York: Wiley.

Blazer, D. G. (1993). *Depression in late life* (2nd ed.). St. Louis, MO: Mosby.

Blind, Dr. (1902). Les debilites mentales. *Revue de Psychiatrie, 8*, 337-345.

Blinkhorn, S. (1988). Lie detection as a psychometric procedure. In A. Gale (ed.), *The polygraph test: Lies, truth and science*. London: SAGE.

Block, J. (1961). *The Q-sort method in personality assessment and psychiatric research*. Springfield, IL: Charles C. Thomas.

Blum, G. (1950). *The Blacky Pictures*. New York: Psychological Corporation.

Blumenthal, J. A. (1985). Review of Jenkins Activity Survey. In: J. V. Mitchell, Jr. (ed.). *The ninth mental measurements yearbook* (vol. 1). Lincoln, NE: Buros Institute of Mental Measurements of the University of Nebraska-Lincoln.

Boggs, D. H., & Simon, J. R. (1968). Differential effect of noise on tasks of varying complexity. *Journal of Applied Psychology, 52*, 148-153.

Bolton, B. (1985). Work Values Inventory. In D. J. Keyser & R. C. Sweetland (eds.), *Test critiques* (vol. 2). Kansas City, MO: Test Corporation of America.

Bond, L. (1989). The effects of special preparation on measures of scholastic ability. In R. L. Linn (ed.), *Educational measurement* (3rd ed.). New York: American Council on Education/Macmillan.

Bondy, M. (1974). Psychiatric antecedents of psychological testing (before Binet). *Journal of the History of the Behavioral Sciences, 10*, 180-194.

Bonner, C. M. (1988). *Utilization of spiritual resources by patients experiencing a recent cancer diagnosis*. Unpublished master's thesis, University of Pittsburgh.

Boring. E. G. (1923, June). Intelligence as the tests test it. *New Republic*, 35-37.

Boring, E. G. (1950). *A history of experimental psychology* (2nd ed). New York: Appleton-Century-Crofts.

Borman, W. C. (1991). Job behavior, performance, and effectiveness. In M. D. Dunnette & L. M. Hough (eds.), *Handbook of industrial and organizational psychology* (vol. 2). Palo Alto, CA: Consulting Psychologists Press.

Borman, W. C., Hanson, M., & Hedge, J. (1997). Personnel selection. *Annual Review of Psychology, 48*, 299-337.

Bornstein, M. H. (1994). Infancy. In R. J. Sternberg (ed.), *Encyclopedia of human intelligence*. New York: Macmillan.

Borstelmann, L. J., & Klopfer, W. G. (1953). The Szondi Tests: A review and critical evaluation. *Psychological Bulletin, 50*, 112-132.

Bouchard, T. J., Jr. (1994). Twin studies. In R. J. Sternberg (ed.), *Encyclopedia of human intelligence*. New York: Macmillan.

Bouchard, T. J., Jr., Lykken, D., McGue, M., Segal, N., & Tellegen, A. (1990). Sources of human psychological differences: The Minnesota Study of Twins Reared Apart. *Science, 250,* 223-228.

Bowers, T., & Pantle, M. (1998). Shipley Institute for Living Scale and the Kaufman Brief Intelligence Test as screening instruments for intelligence. *Assessment, 5,* 187-195.

Boyd, T. M., & Sauter, S. (1993). Route-finding: A measure of everyday executive funtioning in the head-injured adult. *Applied Cognitive Psychology, 7,* 171-181.

Bracken, B. A., and Fagan, T. K. (1990). Guest editors' introduction to the conference "Intelligence: Theories and Practice." *Journal of Psychoeducational Assessment, 8,* 221-222.

Bracken, B. A., Prasse, D. P., & McCallum, R. S. (1984). Peabody Picture Vocabulary Test-Revised: An appraisal and review. *School Psychology Review, 13,* 49-60.

Bradley, R. H., & Caldwell, B. M. (1984). 174 children: A study of the relationship between home environment and cognitive development during the first 5 years. In A. W. Gottfried (ed.), *Home environment and early cognitive development: Longitudinal research.* Orlando, FL: Academic Press.

Bradley, R. H., Mundfrom, D., Whiteside, L., Case, P., & Barrett, K. (1994). A factor analytic study of the Infant-Toddler and Early Childhood versions of the HOME Inventory administered to white, Black, and Hispanic American parents of children born preterm. *Child Development, 65,* 880-888.

Bradley, R. H., & Rock, S. L. (1985). The HOME Inventory: Its relation to school failure and development of an elementary-age version. In W. K. Frankenburg, R. N. Emde, & J. W. Sullivan (eds.), *Early identification of children at risk.* New York: Plenum.

Bradley, R. H., Rock, S. L., Caldwell, B. M., & Brisby, J. A. (1989). Use of the HOME Inventory for families with handicapped children. *American Journal on Mental Retardation, 94,* 313-330.

Bradshaw, J. L., & Mattingley, J. B. (1995). *Clinical neuropsychology: Behavioral and brain science.* San Diego, CA: Academic Press.

Braithwaite, V., & Law, H. (1985). Structure of human values: Testing the adequacy of the Rokeach Value Survey. *Journal of Personality and Social Psychology, 49,* 250-263.

Brannick, M. T., Michaels, C. E., & Baker, D. P. (1989). Construct validity of in-basket scores. *Journal of Applied Psychology, 74,* 957-963.

Brannon, L., & Feist, J. (1992). *Health psychology: An introduction to behavior and health* (2nd ed.). Belmont, CA: Wadsworth.

Brass, D. J., & Oldham, G. R. (1976). Validating an inbasket test using an alternative set of leadership scoring dimensions. *Journal of Applied Psychology, 61,* 652-657.

Breland, H. M. (1979). *Population validity and college entrance measures.* Research Monograph No. 8. New York: The College Board.

Breuer, J., & Freud, S. (1893-1895). Studies on hysteria. In J. Strachey, (ed., in collaboration with A. Freud). *The standard edition of the complete psychological works of Sigmund Freud.* London: Hogarth, 1955, vol. 2.

Bricklin, B. (1992). Data-based tests in custody evaluations. *American Journal of Family Therapy, 20,* 254-265.

Bricklin, B. (1995). *The custody evaluation handbook: Research-based solutions and applications.* New York: Brunner/Mazel.

Brief, D. E., & Comrey, A. L. (1993). A profile of personality for a Russian sample: As indicated by the Comrey Personality Scales. *Journal of Personality Assessment, 60,* 267-284.

Brigham, C. C. (1923). *A study of American intelligence.* Princeton, NJ: Princeton University Press.

Brigham, C. C. (1930). Intelligence tests of immigrant groups. *Psychological Review, 37,* 158-165.

Brinckerhoff, L., Shaw, S., & McGuire, J. (1993). *Promoting Postsecondary Education for Students with Learning Disabilities: A Handbook for Practitioners.* Austin, TX: PRO-ED.

Brodal, A. (1981). *Neurological anatomy* (3rd ed.). New York: Oxford University Press.

Brodsky, S. L. (1989). Advocacy in the guise of scientific advocacy: An examination of Faust and Ziskin. *Computers in Human Behavior, 5,* 261-264.

Brody, E. B., and Brody, N. (1976). *Intelligence: Nature, determinants and consequences.* New York: Academic Press.

Bromberg, W. (1959). *The mind of man: A history of psychotherapy and psychoanalysis.* New York: Harper & Row.

Brooks-Gunn, J., Klebanov, P., & Duncan, G. (1996). Ethnic differences in children's intelligence test scores: Role of economic deprivation, home environment, and maternal characteristics. *Child Development, 67,* 396-408.

Brown, F. G. (1992). Review of Stanford Achievement Test. *Eleventh Mental Measurements Yearbook.* Lincoln, NE: University of Nebraska Press.

Brown, L., Sherbenou, R., & Johnsen, S. (1998). *Test of Nonverbal Intelligence-3.* Austin, TX: Pro-Ed.

Bruininks, R. H. (1978). *Bruininks-Oseretsky Test of Motor Proficiency.* Circle Pines, MN: American Guidance Service.

Bruininks, R. H., Woodcock, R. W., Weatherman, R. F., & Hill, B. K. (1996). *Scales of Independent Behavior-Revised, Interviewer's Manual.* Allen, TX: DLM Teaching Resources.

Bruyere, S. M., & O'Keeffe, J. (eds.). (1994). *Implications of the Americans with Disabilities Act for psychology.* New York: Springer.

Buchsbaum, M. S., & Sostek, A. J. (1980). An adaptiverate continuous performance test: Vigilance characteristics and reliability for 400 male students. *Perceptual and Motor Skills, 51,* 707-713.

Buck, J. (1948). The H-T-P technique, a qualitative and quantitative scoring method. *Journal of Clinical Psychology Monograph Supplement* No. 5, 1-120.

Buck, J. (1981). *The House-Tree-Person technique: A revised manual.* Los Angeles: Western Psychological Services.

Bufford, R., & Parker, T., Jr. (1985). *Religion and wellbeing: Concurrent validation of the Spiritual Well-Being Scale.* Paper presented at the annual meeting of the American Psychological Association, Los Angeles.

Bufford, R., Paloutzian, R., & Ellison, C. (1991). Norms for the Spiritual Well-Being Scale. *Journal of Psychology and Theology, 19,* 56-70.

Burgemeister, B. B., Blum, L. H., & Lorge, I. (1972). *Columbia Mental Maturity Scale* (3rd ed.). New York: Harcourt Brace Jovanovich.

Burke, H. R. (1958). Raven's Progressive Matrices: A review and critical evaluation. *Journal of Genetic Psychology, 93,* 199-228.

Buros, O. K. (1978). *Eighth Mental Measurements Yearbook.* Lincoln, NE: University of Nebraska Press.

Buschke, H., & Fuld, P. A. (1974). Evaluating storage, retention, and retrieval in disordered memory and learning. *Neurology, 24,* 1019-1025.

Bushell, D., Jr. (1978). An engineering approach to the elementary classroom: The Behavior Analysis Follow Through project. In A. C. Catania & T. A. Brigham (eds.), *Handbook of applied behaviour analysis: Social and instructional processes.* New York: Irvington.

Buss, A. (1997). Evolutionary perspectives on personality traits. In R. Hogan, J. Johnson, & S. Briggs (eds.), *Handbook of personality psychology.* San Diego, CA: Academic Press.

Butcher, J. N. (1985). Introduction to the special series. *Journal of Consulting and Clinical Psychology, 53,* 746-747.

Butcher, J. N. (1993). *The Minnesota Report user's guide.* Minneapolis, MN: National Computer System.

Butcher, J. N. (ed.). (1987). *Computerized psychological assessment: A practitioner's guide.* New York: Basic Books.

Butcher, J. N., & Williams, C. L. (1992). *Essentials of MMPI-2 and MMPI-A interpretation.* Minneapolis, MN: University of Minnesota Press.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *Minnesota Multiphasic Personality Inventory-2: Manual for administration and scoring.* Minneapolis, MN: University of Minnesota Press.

Butcher, J. N., Graham, J. R., Williams, C. L., & Ben-Porath, Y. S. (1990). *Development and use of the MMPI-2 content scales.* Minneapolis, MN: University of Minnesota Press.

Caldwell, B. M., & Bradley, R. H. (1984). *Home Observation for Measurement of the Environment.* Little Rock, AR: University of Arkansas at Little Rock.

Caldwell, B. M., & Bradley, R. H. (1994). Environmental issues in developmental follow-up research. In S. L. Friedman & H. C. Haywood (eds.), *Developmental follow-up: Concepts, domains, and methods.* San Diego, CA: Academic Press.

Caldwell, B. M., & Richmond, J. (1967). Social class level and the stimulation potential of the home. In J. Hellmuth (ed.), *The exceptional infant* (vol. 1). Seattle, WA: Special Child Publications.

Callahan, L. A., McGreevy, M., Cirincione, C., and Steadman, H. (1992). Measuring the effects of the Guilty But Mentally Ill (GBMI) verdict. *Law and Human Behavior, 16,* 447-462.

Camara, W. J., & Schneider, D. L. (1994). Integrity tests: Facts and unresolved issues. *American Psychologist, 49,* 112-119.

Campbell, C. D. (1988). Coping with hemodialysis: Cognitive appraisals, coping behaviors, spiritual wellbeing, assertiveness, and family adaptability and cohesion as correlates of adjustment (Doctoral dissertation, Western Conservative Baptist Seminary, 1983). *Dissertation Abstracts International, 49,* 538B.

Campbell, D. P. (1971). *Handbook for the Strong Vocational Interest Blank.* Stanford, CA: Stanford University Press.

Campbell, D. P. (1974). *Manual for the Strong-Campbell Vocational Interest Blank.* Stanford, CA: Stanford University Press.

Campbell, D. P., & Hansen, J. C. (1981). *Manual for the SVIB-SCII* (3rd ed.). Stanford, CA: Stanford University Press.

Campbell, D. P., Hyne, S., & Nilsen, D. (1992). *Manual for the Campbell Interest and Skill Survey.* Minneapolis, MN: National Computer Systems.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

Campbell, J., & McCord, D. (1996). The WAIS-R Comprehension and Picture Arrangement Subtests as measures of social intelligence: Testing traditional interpretations. *Journal of Psychoeducational Assessment, 14*, 240-249.

Campbell, J. P., Dunnette, M. D., Arvey, R. D., & Hellervik, L. V. (1973). The development and evaluation of behaviorally based rating scales. *Journal of Applied Psychology, 57*, 15-22.

Campbell, J. P., Gasser, M., & Oswald, F. (1996). The substantive nature of job performance variability. In K. R. Murphy (ed.), *Individual differences and behavior in organizations*. San Francisco: Jossey-Bass.

Campion, J. E. (1972). Work sampling for personnel selection. *Journal of Applied Psychology, 56*, 40-44.

Campion, M. A., Pursell, E. D., & Brown, B. K. (1988). Structured interviewing: Raising the psychometric properties of the employment interview. *Personnel Psychology, 41*, 25-42.

Canfield, A. A. (1951). The "sten" scale—A modified C-scale. *Educational and Psychological Measurement, 11*, 295-297.

Canivez, G. L. (1995). Validity of the Kaufman Brief Intelligence Test: Comparisons with the Wechsler Intelligence Scale for Children-Third Edition. *Assessment, 2*, 101-111.

Cannell, J. J. (1988). Nationally normed elementary achievement testing in America's public schools: How all 50 states are above the national average. *Educational Measurement: Issues and Practice, 7*, 5-9.

Cantwell, D. P. (ed.). (1975). *The hyperactive child: Diagnosis, management, current research*. New York: Spectrum.

Carlson, C. F., Kula, M., & St. Laurent, C. (1997). Rorschach revised DEPI and CDI with inpatient major depressives and borderline personality disorder with major depression: Validity issues. *Journal of Clinical Psychology, 53*, 51-58.

Carlson, J. S., & Jensen, C. M. (1980). The factorial structure of the Raven Coloured Progressive Matrices Test: A reanalysis. *Educational and Psychological Measurement, 40*, 1111-1116.

Carpenter, C. D. (1992). Review of Stanford Early School Achievement Test. *The Eleventh Mental Measurements Yearbook*. Lincoln, NE: University of Nebraska Press.

Carpenter, M. B. (1991). *Core text of neuroanatomy* (4th ed.). Baltimore, MD: Williams & Wilkins.

Carroll, D. (1988). How accurate is polygraph lie detection? In A. Gale (ed.), *The polygraph test: Lies, truth and science*. London: Sage Publications.

Cascio, W. F. (1976). Turnover, biographical data, and fair employment practice. *Journal of Applied Psychology, 61*, 576-580.

Cascio, W. F. (1987). *Applied psychology in personnel management* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Castlebury, F., Hilsenroth, M., Handler, L., & Durham, T. (1997). Use of the MMPI-2 Personality Disorder Scales in the assessment of DSM-IV Antisocial, Borderline, and Narcissistic Personality Disorders. *Assessment, 4*, 155-168.

Cattell, J. McK. (1890). Mental tests and measurements. *Mind, 15*, 373-380.

Cattell, R. (1950). *Personality: A systematic theoretical and factual study*. New York: McGraw-Hill.

Cattell, R. B. (1940). A culture free intelligence test, Part I. *Journal of Educational Psychology, 31*, 161-179.

Cattell, R. B. (1941). Some theoretical issues in adult intelligence testing. *Psychological Bulletin, 38*, 592 (abstract).

Cattell, R. B. (1971). *Abilities: Their structure, growth, and action*. Boston: Houghton Mifflin.

Cattell, R. B. (1973). Personality pinned down. *Psychology Today, 7*, 40-46.

Cattell, R. B. (1986). *The handbook for the 16 Personality Factor Questionnaire*. Champaign, IL: Institute for Personality and Ability Testing.

Cattell, R. B., & Nesselroade, J. R. (1967). Likeness and completeness theories examined by Sixteen Personality factor measures on stably and unstably married couples. *Journal of Personality and Social Psychology, 7*, 351-361.

Cattell, R. B., Eber, H. W., & Tatsuoka, M. M. (1970). *Handbook for the Sixteen Personality Factor Questionnaire*. Champaign, IL: Institue for Personality and Ability Testing.

Cautela, J. R. (1977). *Behavioral analysis forms for clinical intervention*. Champaign, IL: Research Press.

Ceci, S. J. (1994). Bioecological theory of intellectual development. In R. J. Sternberg (ed.), *Encyclopedia of human intelligence*. New York: Macmillan.

Chaffee, J. W. (1985). *The thorny gates of learning in Sung China: A social history of examinations*. Cambridge: Cambridge University Press.

Chalfant, J. C. (1989). Learning disabilities: Policy issues and promising approaches. *American Psychologist, 44*, 392-398.

Chaplin, W. F. (1984). State-Trait Anxiety Inventory. In: D. J. Keyser & R. C. Sweetland (eds.). *Test critiques* (vol. 1). Kansas City, MO: Test Corporation of America.

Chapman, L. J., & Chapman, J. P. (1967). Genesis of popular but erroneous psychodiagnostic observations. *Journal of Abnormal Psychology, 74*, 271-280.

Choca, J. P., Shanley, L. A., Peterson, C. A., & Van Denburg, E. (1990). Racial bias and the MCMI. *Journal of Personality Assessment, 54*, 479-490.

Choi, H., & Proctor, T. (1994). Error-prone subtests and error types in the administration of the Stanford-Binet Intelligence Scale: Fourth Edition. *Journal of Psychoeducational Assessment, 12*, 165-171.

Cicerelli, V. G. (1969). *The impact of Head Start.* Athens, OH: Westinghouse Learning Corporation and Ohio University.

Clark, D. A. (1988). The validity of measures of cognition: A review of the literature. *Cognitive Therapy and Research, 12*, 1-20.

Clarke, J. M. (1994). Neuroanatomy: Brain structure and function. In D. W. Zaidel (ed.), *Neuropsychology* (2nd ed.). San Diego, CA: Academic Press.

Clarke-Stewart, A., & Friedman, S. (1987). *Child development: Infancy through adolescence.* New York: Wiley.

Clarkin, J. F., Hull, J., Cantor, J., & Sanderson, C. (1993). Borderline personality disorder and personality traits: A comparison of SCID-II BPD and NEO-PI. *Psychological Assessment, 5*, 472-476.

Clarren, S. K., & Smith, D. W. (1978). The fetal alcohol syndrome. *New England Journal of Medicine, 298*, 1063-1067.

Cleary, T. A., Humphreys, L. G., Kendrick, S. A., & Wesman, A. (1975). Educational uses of tests with disadvantage students. *American Psychologist, 30*, 15-41.

Cleckley, H. (1976). *The mask of sanity* (5th ed.). St. Louis, MO: Mosby.

Clemans, W. V. (1971). Test administration. In R. L. Thorndike (ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.

Cleveland, J. N., Murphy, K. R., & Williams, R. E. (1989). Multiple uses of performance appraisal: Prevalence and correlates. *Journal of Applied Psychology, 74*, 130-135.

Cohen, M. S., & Bookheimer, S. (1994). Localization of brain function using magnetic resonance imaging. *Trends in Neurosciences, 17*, 268-277.

Colby, A., & Kohlberg, L. (1987). *The measurement of moral judgment* (vol. I). Cambridge: Cambridge University Press.

Colby, A., Kohlberg, L., Gibbs, J. C., & others. (1978). *Measuring moral judgment: Standardized scoring manual.* Cambridge, MA: Harvard University, Moral Education Research Foundation.

Colby, A., Kohlberg, L., Gibbs, J., & Lieberman, M. (1983). A longitudinal study of moral judgment. *Monographs for the Society for Research in Child Development, 48*, 1, 2.

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In Linn, R. L. (ed.), *Educational measurement* (3rd ed.). New York: ACE/Macmillan.

Collins, J. M., & Schmidt, F. L. (1993). Personality, integrity, and white collar crime: A construct validity study. *Personnel Psychology, 46*, 295-311.

Committee on Ethical Guidelines for Forensic Psychologist. (1991). Specialty guidelines for forensic psychologist. *Law and Human Behavior, 15*, 655-665.

Comrey, A. L. (1970). *Manual for the Comrey Personality Scales.* San Diego, CA: EdITS.

Comrey, A. L. (1973). *A firts course in factor analysis.* New York: Academic Press.

Comrey, A. L. (1980). *Handbook of interpretations for the Comrey Personality Scales.* San Diego, CA: EdITS.

Comrey, A. L., & Backer, T. (1970). Construct validation of the Comrey Personality Scales. *Multivariate Behavior Research, 5*, 469-477.

Comrey, A. L., & Schiebel, D. (1983). Personality test correlates of psychiatric outpatient status. *Journal of Consulting and Clinical Psychology, 51*, 756-762.

Comrey, A. L., & Schiebel, D. (1985). Personality test correlates of psychiatric case history data. *Journal of Consulting and Clinical Psychology, 53*, 470-479.

Conners, C. K. (1990). *Conners' Rating Scales.* Los Angeles: Western Psychological Services.

Conners, C. K. (1991). *Conners' Teacher Rating Scales-39.* North Tonawanda, NY: Multi-Health Systems, Inc.

Conoley, C. W. (1992). Review of Beck Depression Inventory. *The Eleventh Mental Measurements Yearbook.* Lincoln, NE: University of Nebraska Press.

Conoley, C. W., Plake, B., and Kemmerer, B. (1991). Issues in computer-based test interpretive systems. *Computers in Human Behavior, 7*, 97-101.

Conoley, J. C., & Impara, J. C. (eds.). (1995). *The Twelfth Mental Measurements Yearbook.* Lincoln, NE: The University of Nebraska Press.

Conoley, J. C., & Kramer, J. J. (eds.). (1989). *The Tenth Mental Measurements Yearbook.* Lincoln, NE: The Buros Institute of Mental Measurements.

Conoley, J. C., & Kramer, J. J. (eds.). (1992). *The Eleventh Mental Measurements Yearbook.* Lincoln, NE: The Buros Institute of Mental Measurements.

Conry, R., & Plant, W. T. (1965). WAIS and group test predictions of an academic success criterion: High School and college. *Educational and Psychological Measurement, 25*, 493-500.

Constantino, G., & Malgady, R. (1996). Development of TEMAS, a multicultural thematic apperception test: Psychometric properties and clinical utility. In G. R. Sodowsky & J. C. Impara (eds.), *Multicultural assessment in counseling and clinical psychology.* Lincoln, NE: The Buros Institute of Mental Measurements.

Constantino, G., Malgady, R., & Rogler, L. (1988). *Tell-Me-A-Story (TEMAS): manual.* Los Angeles: Western Psychological Services.

Conway, J. M., Jako, R., & Goodman, D. (1995). A metaanalysis of interrater and internal consistency reliability of selection interviews. *Journal of Applied Psychology, 80,* 565-579.

Cooley, E. J., & Ayres, R. (1985). Convergent and discriminant validity of the Mental Processing Scales of the Kaufman Assessment Battery for Children. *Psychology in the Schools, 22,* 373-377.

Cooper, D. (1990). Factor structure of the Edwards Personal Preference Schedule in a vocational rehabilitation sample. *Journal of Clinical Psychology, 46,* 421-424.

Cooper, D., & Shepard, K. (1992). Review of DIAL-R. *Learning Disabilities Research & Practice, 7,* 171-174.

Corcoran, K., & Fischer, J. (1994). *Measures for clinical practice: A sourcebook, Vol. 1,* 2. (2nd ed.). New York: Free Press.

Cornelius, E. T., & Hakel, M. D. (1978). *A study to develop and improved enlisted performance evaluation system for the U.S. Coast Guard.* Washington, DC: Department of Transportation, USCG.

Cornelius, S. W., and Caspi, A. (1987). Everyday Problem Solving in Adulthood and Old Age. *Psychology and Aging, 2,* 144-153.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78,* 98-104.

Cosden, M. (1992). Review of the Draw A Person: A Quantitative Scoring System. *Them Eleventh Mental Measurements Yearbook.* Lincoln, NE: University of Nebraska Press.

Costa, P. T., Jr. (1991). Clinical use of the five-factor model. *Journal of Personality Assessment, 57,* 393-398.

Costa, P. T., Jr., & McCrae, R. (1989). *NEO Five-Factor Inventory test manual.* Port Huron, MI: Sigma Assessment Systems.

Costa, P. T., Jr., & McCrae, R. (1992). *NEO PI-R test manual.* Port Huron, MI: Sigma Assessment Systems.

Costa, P. T., Jr., McCrae, R. R., & Holland, J. L. (1984). Personality and vocational interests in an adult sample. *Journal of Applied Psychology, 69,* 390-400.

Cote, L., & Crutcher, M. D. (1991). The basal ganglia. In E. R. Kandel, J. H. Schwartz, & T. M. Jessell (eds.), *Principles of neural science* (3rd ed.). New York: Elsevier.

Coveny, T. E. (1972). A new test for the visually handicapped: Preliminary analysis o reliability and validity of the Perkins-Binet. *Education of the Handicapped, 4,* 97-101.

Cowdery, K. M. (1926-27). Measurement of professional attitudes: Differences between lawyers, physicians, and engineers. *Journal of Personnel Research, 5,* 131-141.

Craig, R. J. (ed.). (1993). *The Millon Clinical Multiaxial Inventory: A clinical research information synthesis.* Hillsdale, NJ: Erlbaum.

Crandall, J. E. (1981). *Theory and measurement of social interest: Empirical tests of Alfred Adler's concept.* New York: Columbia University Press.

Crandall, J. E. (1984). Social interest as a moderator of life stress. *Journal of Personality and Social Psychology, 47,* 164-174.

Crapo, L. (1985). *Hormones: The messengers of life.* New York: Freeman.

Crawford, J. R., Sommerville, J., & Robertson, I. (1997). Assessing the reliability and abnormality of subtest differences on the Test of Everyday Attention. *British Journal of Clinical Psychology, 36,* 609-617.

Cripe, L. (1996). The ecological validity of executive function testing. In R. J. Sbordone & C. J. Long (eds.), *Ecological validity of neuropsychological testing.* Delray Beach, FL: GR Press/St. Lucie Press.

Critchley, M. (1953). *The parietal lobes.* London: Edward Arnold.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334.

Cronbach, L. J. (1970). *Essentials of psychological testing* (3rd ed.). New York: Harper & Row.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (ed.), *Educational Measurement* (2nd ed.). Washington, DC: American Council on Education.

Cronbach, L. J. (1984). *Essentials of psychological testing* (4th ed.). New York: Harper & Row.

Cronbach, L. J. (1988). Five perspectives on the validity argument. In H. Wainer & H. I. Braun (eds.), *Test validity,* Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions.* Urbana, IL: University of Illinois Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Crosson, B., & Warren, R. L. (1982). Use of the Luria-Nebraska Neuropsychological Battery. *Journal of Consulting and Clinical Psychology, 50,* 22-31.

Culbertson, J., & Edmonds, A. (1996). Learning Disabilities. In R. Adams, O. Parsons, J. Culbertson, & S. Nixon, (eds.). *Neuropsychology for*

clinical practice: Etiology, assessment, and treatment of common neurological disorders. Washington, DC: American Psychological Association.

Cunningham, J. W., & Ballentine, R. D. (1982). *The General Work Inventory*. Raleigh, NC: Authors.

Cunningham, J. W., Boese, R, Neeb, R., & Pass, J. (1983). Systematically derived work dimensions: Factor analyses of the Occupation Analysis Inventory. *Journal of Applied Psychology, 68*, 232-252.

Cunningham, M., Wong, D., & Barbee, A. (1994). Self-presentation dynamics on overt integrity tests: Experimental studies with the Reid Report. *Journal of Applied Psychology, 79*, 643-658.

Cureton, E. E. (1950). Validity, reliability, and baloney. *Educational and Psychological Measurement, 10*, 94-96.

Cytowic, R. E. (1996). *The neurological side of neuropsychology*. Cambridge, MA: MIT Press.

Dahlstrom, W. G., Welsh, G. S., & Dahlstrom, L. E. (1972). *An MMPI handbook. Volume I: Clinical interpretation* (rev.ed.). Minneapolis, MN: University of Minnesota Press.

Dahlstrom, W. G., Welsh, G. S., & Dahlstrom, L. E. (1975). *An MMPI handbook: Vol. II. Research applications*. Minneapolis, MN: University of Minnesota Press.

Dana, R. H. (1959). Proposal for objective scoring of the TAT. *Perceptual and Motor Skills, 10*, 27-43.

Dana, R. H. (1993). *Multicultural assessment perspectives for professional psychology*. Boston: Allyn & Bacon.

Dana, R. H., & Cantrell, J. D. (1988). An update on the Millon Clinical Multiaxial Inventory (MCMI). *Journal of Clinical Psychology, 44*, 760-763.

Darlington, R. B. (1964). Increasing test validity through the use of interitem correlations (doctoral dissertation, University of Minnesota, 1963). *Dissertation Abstracts International, 24*, 4778. (University Microfilms No. 64-4089).

Darlington, R. B., & Bishop, C. H. (1966). Increasing test validity by considering interitem correlations, *Journal of Applied Psychology, 50*, 322-330.

Das, J. P. (1994). Serial and parallel processing. In R. J. Sternberg (ed.), *Encyclopedia of human intelligence*. New York: Macmillan.

Das, J. P., and Mensink, D. (1989). K-ABC Simultaneous-Sequential Scales and prediction of achievement in reading and mathematics. *Journal of Psychoeducational Assessment, 7*, 103-111.

Das, J. P., and Naglieri, J. A. (1993). *Cognitive assessment system: Standardization version*. Chicago, IL: Riverside.

Das, J. P., Naglieri, J., & Kirby, J. (1994). *Assessment of cognitive processes: The PASS theory of intelligence*. Boston: Allyn & Bacon.

Das, J. P., Kirby, J. R., & Jarman, R. F. (1979). *Simultaneous and successive cognitive processes*. Orlando, FL: Academic Press.

Davila, M., Shear, P., Lane, B., Sullivan, E., & Pfefferbaum, A. (1994). Mamillary body and cerebellar shrinkage in chronic alcoholics: An MRI and neuropsychological study. *Neuropsychology, 8*, 433-444.

Davis, C. (1980). *Perkins-Binet Tests of Intelligence for the blind*. Watertown, MA: Perkins School for the Blind.

Davison, M., Gasser, M., & Ding, S. (1996). Identifying major profile patterns in a population: An exploratory study of WAIS and GATB patterns. *Psychological Assessment, 1*, 26-31.

Dawes, R. M., Faust, D., and Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*, 1668-1674.

Dawis, R. V., & Lofquist, L. H. (1984). *A psychological theory of work adjustment: An individual-differences model and its applications*. Minneapolis, MN: University of Minnesota Press.

Dawis, R. V., England, G. W., & Lofquist, L. H. (1964). A theory of work adjustment. *Minnesota Studies in Vocational Rehabilitation*, 15.

Dean, R. S. (1986). Lateralization of cerebral functions. In D. Wedding, A. M. Horton, & J. Webster (eds.), *The neuropsychology handbook*. New York: Springer.

Deary, I. J., Hendrickson, A. E., and Burns, A. (1987). Serum calcium levels in Alzheimer's disease: A finding and an aetiological hypothesis. *Personality and Individual Differences, 8*, 75-80.

DeCrans, M. (1990). *Spiritual well-being in the rural elderly*. Unpublished manuscript, Marquette University, Milwaukee, WI.

Delis, D. C., & Kaplan, E. (1982). Assessment of aphasia with the Luria-Nebraska Neuropsychological Battery: A conceptual critique. *Journal of Consulting and Clinical Psychology, 50*, 32-39.

Delis, D. C., & Kaplan, E. (1983). Hazards of a standardized neuropsychological test with low content validity. *Journal of Consulting and Clinical Psychology, 51*, 396-398.

Dembroski, T., MacDougall, J., Williams, B., & Haney, T. (1985). Components of Type A, hostility, and anger-in: Relationship to angiographic findings. *Psychosomatic Medicine, 47*, 219-233.

Deri, S. (1949). *Introduction to the Szondi Test*. New York: Grune & Stratton.

Detterman, D. K. (1984). Understand cognitive components before postulating metacomponents, etc., part 2. *Behavioral and Brain Sciences, 7*, 289-290.

DeVito, A. J. (1985). Review of Myers-Briggs Type Indicator. *Ninth mental measurements yearbook*. Lincoln, NE: University of Nebraska Press.

Diamond, S. (1980). Wundt before Leipzig. In R. W. Rieber (ed.), *Wilhelm Wundt and the making of a scientific psychology*. New York: Plenum Press.

Digman, J. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology, 41*, 417-440.

Dikmen, S., Machamer, J., Winn, H., & Temkin, N. (1995). Neuropsychological outcome at 1-year post head injury. *Neuropsychology, 9*, 80-90.

DiLalla, L. F., Thompson, L. A., Plomin, R., and others. (1990). Infant predictors of preschool and adult IQ: A study of infant twins and their parents. *Developmental Psychology, 26*, 759-769.

Dillon, H. J. (1949). *Early school leavers: A major educational problem*. New York: National Child Labor Committee.

Dillon, R. F., Pohlmann, J. T., & Lohman, D. F. (1981). A factor analysis of Raven's Advanced Progressive Matrices freed of difficulty factors. *Eduational and Psychological Measurement, 41*, 1295-1302.

Dixon, D. (1995). Review of Structured Interview of Reported Symptoms. *Twelfth Mental Measurements Yearbook*. Lincoln, NE: University of Nebraska Press.

Dodrill, C. B. (1979). Sex differences on the Halstead-Reitan Neuropsychological Battery and on other neuropsychological measures. *Journal of Clinical Psychology, 35*, 236-241.

Dodrill, C. B. (1981). An economical method of measuring general intelligence in adults. *Journal of Consulting and Clinical Psychology, 49*, 668-673.

Dodrill, C. B., & Warner, M. H. (1988). Further studies of the Wonderlic Personnel Test as a brief measure of intelligence. *Journal of Consulting and Clinical Psychology, 56*, 145-147.

Doll, E. A. (1935). The Vineland Social Maturity Scale. *Training School Bulletin, 32*, 1-7, 25-32, 48-55, 68-74.

Doll, E. A. (1936). Preliminary standardization of the Vineland Social Maturity Scale. *American Journal of Orthopsychiatry, 6*, 283-293.

Dolliver, R. H., Irvin, J. A., & Bigley, S. E. (1972). Twelve-year follow-up of the Strong Voctional Interest Blank. *Journal of Counseling Psychology, 19*, 212-217.

Donders, J. (1995). Validity of the Kaufman Brief Intelligence Test (K-BIT) in children with traumatic brain injury. *Assessment, 2*, 219-224.

Donlon, T. F. (ed.). (1984). *The College Board technical handbook for the Scholastic Aptitude Test and Achievement Tests*. New York: College Entrance Examination Board.

Donnay, D., & Borgen, F. (1996). Validity, structure, and content of the 1994 Strong Interest Inventory. *Journal of Counseling Psychology, 43*, 275-291.

Doppelt, W. E. (1956). Estimating the full scale score on the Wechsler Adult Intelligence Scale from scores on four subtests. *Journal of Consulting Psychology, 20*, 63-66.

Drakeley, R. J., Herriot, P., & Jones, A. (1988). Biographical data, training success and turnover. *Journal of Occupational Psychology, 61*, 145-152.

Drebing, C., Van Gorp, W., Stuck, A., Mitrushina, M., & Beck, J. (1994). Early detection of cognitive decline in higher cognitively functioning older adults: Sensitivity and specificity of a neuropsychological screening battery. *Neuropsychology, 8*, 31-37.

Drummond, R. J. (1987). Edwards Personal Preference Schedule. In D. J. Keyser & R. C. Sweetland (eds.), *Test critiques compendium*. Kansas City, MO: Test Corporation of America.

DuBois, P. E. (1939). A test standardized on Pueblo Indian children. *Psychological Bulletin, 36*, 523.

DuBois, P. H. (1970). *A history of psychological testing*. Boston: Allyn & Bacon.

Dumenci, L. (1995). Construct validity of the Self-Directed Search using hierarchically nested structural models. *Journal of Vocational Behavior, 47*, 21-34.

Dumont, R., & Hagberg, C. (1994). Kaufman Adolescent and Adult Intelligence Test (KAIT). *Journal of Psychoeducational Assessment, 12*, 190-196.

Dumont, R., Cruse, C., Price, L., & Whelley, P. (1996). The relationship between the Differential Ability Scales (DAS) and the Wechsler Intelligence Scale for Children-Third Edition (WISC-III). *Psychology in the Schools, 33*, 203-209.

Dunn, L. M., & Dunn, L. M. (1981). *Peabody Picture Vocabulary Test-Revised*. Circle Pines, MN: American Guidance Service.

Dunn, L. M., & Dunn, L. M. (1998). *Examiner's Manual: Peabody Picture Vocabulary Test-III*. Cicle Pines, MN: American Guidance Service.

Dunnette, M. D., & L. M. Hough (ed.). (1990-1992). *Handbook of industrial and organizational psychology* (vols. 1-3). Palo Alto, CA: Consulting Psychologists Press.

Dunst, C. J. (1980). *A clinical and educational manual for use with the Uzgiris and Hunt Scales of Infant Psychological Development*. Baltimore: University Park Press.

Dyce, J. A. (1996). Factor structure of the Beck Hopelessness Scale. *Journal of Clinical Psychology, 52*, 555-558.

Eaker, E. D., & Castelli, W. P. (1988). Type A behavior and coronary heart disease in women: Fourteen-year incidence from the Framingham study. In B. K. Houston & C. R. Snyder (eds.), *Type A behavior pattern: Research, theory, and interventions*. New York: Wiley.

Ebbinghaus, H. (1987). Ueber eine neue Methode zur Pruefung geistiger Faehigkeiten und ihre Anwen-

dung bei Schulkindern. *Zeitschrift fuer Angewandte Psychologie, 13*, 401-459.

Eccles, J. C. (1973). *The understanding of the brain.* New York: McGraw-Hill.

Educational Testing Service. (1987). *ETS standards for quality and fairness.* Princeton, NJ: Author.

Educational Testing Service. (1988). *Guidelines for proper use of NTE tests.* Princeton, NJ: Author.

Educational Testing Service. (1989). *Guidelines for proper use of GRE scores.* Princeton, NJ: Author.

Edwards, A. L. (1957). *The social desirability variable in personality assessment and research.* New York: Dryden Press.

Edwards, A. L. (1959). *Edwards Personal Preference Schedule.* New York: The Psychological Corporation.

Edwards, A. L. (1961). Social desirability or acquiescence in the MMPI? A case study with the SD Scale. *Journal of Abnormal and Social Psychology, 63*, 351-359.

Edwards, A. L. (1967). *Edwards Personality Inventory.* Chicago: Science Research Associates.

Edwards, A. L., & Abbott, R. D. (1973). Relationships among the Edwards Personality Inventory Scales, the Edwards Personality Preference Schedule, and the Personality Research Form Scales. *Journal of Consulting and Clinical Psychology, 40*, 27-32.

Efron, R., (1990). *The decline and fall of hemispheric specialization.* Hillsdale, NJ: Erlbaum.

Eisdorfer, C., & Cohen, L. D. (1961). The generality of the WAIS standardization for the aged: A regional comparison. *Journal of Abnormal and Social Psychology, 62*, 520-527.

Eisenson, J. (1954). *Examining for aphasia.* New York: Psychological Corporation.

Eisenstein, N., & Engelhart, C. (1997). Comparison of the K-BIT with short forms of the WAIS-R in a neuropsychological population. *Psychological Assessment, 9*, 57-62.

Ekman, P. (1982). Methods for measuring facial action. In K. R. Scherer & P. Ekman (eds.), *Handbook of methods in nonverbal behavior research.* Cambridge: Cambridge University Press.

Ekman, P., & Friesen, W. V. (1975). *Unmasking the face.* Englewood Cliffs, NJ: Prentice-Hall.

Ekman, P., & Friesen, W. V. (1978). *The Facial Action Coding System: A technique for the measurement of facial movement.* Palo Alto, CA: Consulting Psychologists Press.

Ekman, P., & Friesen, W. V., & Ancoli, S. (1980). Facial signs of emotional experience. *Journal of Personality and Social Psychology, 39*, 1125-1134.

Elaad, E., Ginton, A., & Jungman, N. (1992). Detection measures in real-life criminal guilty knowledge tests. *Journal of Applied Psychology, 77*, 757-767.

Elliott, C. D. (1990). *The Differential Ability Scales: Introductory and technical handbook.* San Antonio, TX: Psychological Corporation.

Elliott, C. D. (1997). The Differential Ability Scales. In D. P. Flanagan, J. Genshaft, & P. Harrison (eds.), *Contemporary intellectual assessment: Theories, tests, and issues.* New York: Guilford.

Ellis, A. (1962). *Reason and emotion in psychotherapy.* New York: Lyle Stuart.

Ellis, A. (1970). H-T-P: A projective device and a measure of adult intelligence. In O. K. Buros (ed.), *Personality tests and reviews.* Highland Park, NJ: The Gryphon Press.

Ellison, C. W. (1983). Spiritual well-being: Conceptualization and measurement. *Journal of Psychology and Theology, 11*, 330-340.

Ellison, C. W., & Smith, J. (1991). Toward an integrative measure of health and well-being. *Journal of Psychology and Theology, 19*, 35-48.

Erdberg, P. (1985). The Rorschach. In C. S. Newmark (ed.), *Major psychological assessment instruments.* Boston: Allyn & Bacon.

Eron, L. D. (1950). A normative study of the Thematic Apperception Test. *Psychological Monographs, 64* (9,, Whole No. 315).

Ertl, J. P., and Schafer, E. W. P. (1969). Brain response correlates of psychometric intelligence. *Nature, 223*, 421-422.

Esquirol, J. E. D. (1845/1838). *Mental maladies.* (trans. E. K. Hunt). Philadelphia: Lea & Blanchard.

Estes, W. K. (1974). Learning theory and intelligence. *American Psychologist, 29*, 740-749.

Evans, D. A., Funkenstein, H., Albert, M., & others (1989). Prevalence of Alzheimer's Disease in a community population of older persons. *Journal of the American Medical Association, 262*, 2551-2556.

Exline, R. V., & Fehr, B. J. (1982). The assessment of gaze and mutual gaze. In K. R. Scherer & P. Ekman (eds.), *Handbook of methods in nonverbal behavior research.* Cambridge: Cambridge University Press.

Exner, J. E., Jr. (1991). *The Rorschach: A comprehensive system, Volume 2. Current research and advanced interpretation* (2nd ed.). New York: Wiley.

Exner, J. E., Jr. (1993). *The Rorschach: A comprehensive system, Volume 1. Basic foundations* (3rd ed.). New York: Wiley.

Exner, J. E., Jr. (1995). *Issues and methods in Rorschach research.* Mahwah, NJ: Erlbaum.

Exner, J. E., Jr., & Andronikof-Sanglade, A. (1992). Rorschach changes following brief and short-term psychotherapy. *Journal of Personality Assessment, 59*, 59-71.

Exner, J. E., Jr., & Weiner, I. B. (1994). *The Rorschach: A comprehensive system, Volume 3. Assessment of children and adolescents* (2nd ed.). New York: Wiley.

Eyde, L. D., and Primhoff, E. S. (1992). Responsible test use. In M. Zeidner and R. Most (eds.),

Psychological testing: An inside view. Palo Alto, CA: Consulting Psychologists Press.

Eyde, L. D., Robertson, G. J., Krug, S., and others. (1993). *Responsible test use: Case studies for assessing human behavior.* Washington, DC: American Psychological Association.

Eysenck, H. J. (1986). Is intelligence? IN R. J. Sternberg and D. K. Detterman (eds.), *What is intelligence? Contemporary viewpoints on its nature and definition.* Norwood, NJ: Ablex.

Eysenck, H. J. (1994). EEG evoked potentials. In R. J. Sternberg (ed.), *Encyclopedia of human intelligence.* New York: Macmillan.

Eysenck, H. J. (ed.). (1982). *A model for intelligence.* Heidelberg: Springer-Verlag.

Eysenck, H. J., & Eysenck, M. W. (1975). *Manual of the Eysenck Personality Questionnaire.* San Diego: Educational and Industrial Testing Service.

Eysenck, H. J., & Eysenck, M. W. (1985). *Personality and individual differences: A natural science approach.* New York: Plenum.

Eysenck, H. J., & Eysenck, S. B. G. (1976). *Psychoticism as a dimension of personality.* London: Hodder & Stoughton.

Fagan, J. F. III. (1984). Infant memory. In M. Moscovitch (ed.), *Infant memory.* New York: Plenum Press.

Fagan, J. F. III, & Haiken-Vasen, J. (1997). Selective attention to novelty as a mesaure of information processing across the lifespan. In J. Burack & J. Enns (eds.), *Attention, development, and psychopathology.* New York: Guilford.

Fagan, J. F. III, & McGrath, S. K. (1981). Infant recognition memory and later intelligence. *Intelligence, 5,* 121-130.

Fagan, J. F. III. & Shepherd, P.A. (1986). *The Fagan Test of Infant Intelligence: Training manual.* Cleveland, OH: Infantest Corporation.

Fancher, R. E. (1985). *The intelligence men: Makers of the IQ controversy.* New York: Norton.

Farnham-Diggory, S. (1978). *Learning disabilities: A psychological perspective.* Cambridge, MA: Harvard University Press.

Faust, D., and Ziskin, J. (1988). The expert witness in psychology and psychiatry. *Science, 241,* 31-35.

Faust, D., and Ziskin, J. (1989). Computer-assisted psychological evaluation as legal evidence: Some day my prints will come. *Computers in Human Behavior, 5,* 23-36.

Federal Rules of Evidence for United States Courts and Magistrates. (1975). St. Paul, MN: West Publishing Company.

Fehr, B. J., & Exline, R. V. (1987). Social visual interaction: A conceptual and literature review. In A. W. Siegman & S. Feldstein (eds.), *Nonverbal behavior and communication* (2nd ed.). Hillsdale, NJ: Erlbaum.

Fehring, R., Brennan, P., & Keller, M. (1987). Psychological and spiritual well-being in college students. *Research in Nursing and Health, 10,* 391-398.

Feldman, R. D. (1982). *Whatever happened to the quiz Kids?* Chicago: Chicago Review Press.

Feldt, L. S., & Brennan, R. L. (1989). In R. L. Linn (ed.), *Educational Measurement* (3rd ed.). New York: American Council on Education/Macmillan.

Ferris, G., Judge, T., Rowland, K., & Fitzgibbons, D. (1994). Subordinate influence and the performance evaluation process: Test of a model. *Organizational Behavior and Human Decision Processes, 58,* 101-135.

Ferris, S. H. (1992). Diagnosis by specialists: Psychological testing. *Acta Neurologica Scandinavica, 85,* 32-35.

Figueroa, R. A. (1990). Best practices in the assessment of bilingual children. In A. Thomas & J. Grimes (eds.), *Best practices in school psychology II.* Washington, DC: National Association of School Psychologists.

Finn, S. E. (1996). *A manual for using the MMPI-2 as a therapeutic intervention.* Minneapolis: University of Minnesota Press.

Finn, S. E., & Tonsager, M. E. (1992). Therapeutic effects of providing MMPI-2 test feedback to college students awaiting therapy. *Psychological Assessment, 4,* 278-287.

Finn, S. E., & Tonsager, M. E. (1997). Information-gathering and therapeutic models of assessment: Complementary paradigms. *Psychological Assessment, 9,* 374-385.

Fisher, G. L., Jenkins, S. J., Bancroft, M. J., and Kraft, L. M. (1988). The effects of K-ABC-based remedial teaching strategies on word recognition skills. *Journal of Learning disabilities, 21,* 307-312.

Fisher, S., & Greenberg, R. P. (1984). *The scientific credibility of Freud's theories and therapy.* New York: Columbia University Press.

Fiske, D. W. (1986). The trait concept and the personality questionnaire. In A. Angleitner & J. S. Wiggins (eds.), *Personality assessment via questionnaires: Current issues in theory and measurement.* Berlin: Springer-Verlag.

Fitzpatrick, R. (1984). Employee Aptitude Survey Tests. In D. J. Keyser & R. C. Sweetland (eds.), *Test Critiques* (vol. 1). Kansas City, MO: Test Corporation of America.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51,* 327-358.

Flanagan, J. C. (1956). The evaluation of methods in applied psychology and the problem of criteria. *Occupational Psychology, 30,* 1-9.

Fleishman, E. A. (1975). Toward a taxonomy of human performance. *American Psychologist, 30,* 1127-1149.

Fleishman, E. A., & Mumford, M. D. (1988). Ability requirement scales. In S. Gael (ed.), *Job analysis handbook for business, industry, and government* (vol. 2). New York: Wiley.

Fleishman, E. A., & Quaintance, M. K. (1984). *Taxonomies of human performance*. New York: Academic Press.

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin, 95*, 29-51.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101*, 171-191.

Flynn, J. R. (1994). IQ gains over time. In R. J. Sternberg (ed.), *Encyclopedia of human intelligence*. New York: Macmillan.

Folstein, M., Folstein, S., & McHugh, P. (1975). Mini-Mental State: A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 12*, 189-198.

Forns-Santacana, M., and Gomez-Benito, J. (1990). Factor structure of the McCarthy Scales. *Psychology in the Schools, 27*, 111-115.

Forrest, D. W. (1974). *Francis Galton: The life and work of a Victorian genius*. New York: Taplinger Publishing.

Forster, A. A. (1994). Learning Disability. In R. J. Sternberg (ed.), *Encyclopedia of human intelligence*. New York: Macmillan.

Foster, C. (1990). Achievement tests put pressure on students and schools. *Cristian Science Monitor*, May 24, 7.

Foster, R. W. (1974). *Camelot Behavioral Checklist manual*. Bellevue, WA: Edmark Associates.

Foster, S. L., Bell-Dolan, D. J., & Burge, D. A. (1988). Behavioral observation. In A. S. Bellack & M. Hersen (eds.), *Behavioral assessment: A practical handbook* (3rd ed.). New York: Pergamon Press.

Fowler, J. (1981). *Stages of faith: The psychology of human development and the quest for meaning*. San Francisco: Harper & Row.

Fowler, P. C., Zillmer, E., & Macciocchi, S. N. (1990). Confirmatory factor analytic models of the WAIS-R for neuropsychiatric patients. *Journal of Clinical Psychology, 46*, 324-332.

Fowler, R. D. (1985). Landmarks in computer-assisted psychological assessment. *Journal of Consulting and Clinical Psychology, 53*, 748-759.

Fowler, R. D., & Matarazzo, J. D. (1988). Psychologists and psychiatrists as expert witnesses. *Science, 241*, 1143.

Frank, G. (1983). *The Wechsler enterprise: An assessment of the development, structure, and use of the Wechsler tests of intelligence*. New York: Pergamon Press.

Frank, G. (1990). Research on the clinical usefulness of the Rorschach: 1. The diagnosis of schizophrenia. *Perceptual and Motor Skills, 71*, 573-578.

Frank, L. K. (1939). Projective methods for the study of personality. *Journal of Psychology, 8*, 389-413.

Frank, L. K. (1948). *Projective methods*. Springfield, IL: Thomas.

Franke, W. (1963). *The reform and abolition of the traditional Chinese examination system*. Cambridge, MA: Harvard University Press.

Frankenburg, W. K. (1985). The Denver approach to early case finding: A review of the Denver Developmental Screening Test and a brief training program in developmental diagnosis. In W. K. Frankenburg, R. M. Emde, and J. W. Sullivan (eds.), *Identification of children at risk: An international perspective*. New York: Plenum Press.

Frankenburg, W. K., and Dodds, J. B. (1967). The Denver developmental screening tests. *Journal of Pediatrics, 71*, 181-191.

Frankenburg, W. K., Dodds, J., Archer, P., and others. (1990). *Denver II: Technical manual*. Denver, CO: Denver Developmental Materials.

Frankl, V. (1963). *Man's search for meaning: An introduction to logotherapy*. New York: Washington Square Press.

Frauenheim, J. G., & Heckerl, J. R. (1983). A longitudinal study of psychological and achievement test performance in severe dyslexic adults. *Journal of Learning Disabilities, 16*, 339-347.

Frechtling, J. A. (1989). Administrative uses of school testing programs. In R. L. Linn (ed.), *Educational measurement* (3rd ed.). New York: American Council on Education/Macmillan.

Frederickson, L. C. (1985). Goodenough-Harris Drawing Test. In Keyser, D. J., & Sweetland, R. C. (eds.). *Test critiques*(vol. 2). Kansas City, MO: Test Corporation of America.

Frederiksen, N. (1962). Factors in In-basket Performance. *Psychological Monographs, 76*, Whole No. 541.

Freud, A. (1946). *The ego and the mechanisms of defense*. New York: International Universities Press.

Freud, S. (1900). The interpretation of dreams. In J. Strachey, (ed., in collaboration with A. Freud). *The standard edition of the complete psychological works of Sigmund Freud*. London: Hogarth, 1955, vols. 4 and 5.

Freud, S. (1927/1961). *The future of an illusion* (J. Strachey, trans.) New York: Basic Books. (Originally published 1900).

Freud, S. (1933). *New introductory lectures on psychoanalysis*. New York: Norton.

Fridlund, A. J., Ekman, P., & Oster, H. (1987). Facial expressions of emotion: Review of literature 1970-1983. In A. W. Siegman & S. Feldstein (eds.), *Nonverbal behavior and communication* (2nd ed.). Hillsdale, NJ: Erlbaum.

Friedman, A. F. (1987). Eysenck Personality Questionnaire. In D. J. Keyser & R. C. Sweetland (eds.), *Test critiques compendium*. Kansas City, MO: Test Corporation of America.

Friedman, M., & Rosenman, R. (1974). *Type A behavior and your heart*. New York: Fawcett Crest.

Friedman, M., & Ulmer, D. (1984). *Treating Type A behavior and your heart*. New York: Alfred A. Knopf.

Fuess, C. M. (1950). *The College Board: Its first fifty years*. New York: Columbia University Press.

Fuld, P. A. (1977). *Fuld Object-Memory Evaluation*. Chicago: Stoelting Co.

Fuld, P. A., Masur, D. M., Blau, A. D., Crystal, H., & Aronson, M. K. (1990). Object-Memory Evaluation for prospective detection of dementia in normal functioning elderly: predictive and normative data. *Journal of Clinical and Experimental Neuropsychology, 12*, 520-528.

Fuller, G. B., Parmelee, W. M., & Carroll, J. L. (1982). Performance of delinquent and non-delinquent high school boys on the Rotter Incomplete Sentences Blank. *Journal of Personality Assessment, 46*, 506-510.

Furnham, A., Toop, A., Lewis, C., & Fisher, A. (1995). P-E fit and job satisfaction: A failure to support Holland's theory in three British samples. *Personality and Individual Differences, 19*, 677-690.

Gale, A., and Edwards, J. (1983). Cortical correlates of intelligence. In A. Gale and J. Edwards (eds.), *Physiological correlates of human behaviour* (vol. 3). New York: Academic Press.

Gallagher, J. J., & Courtright, R. D. (1986). The educational definition of giftedness and its policy implications. In R. J. Sternberg and J. E. Davidson (eds.), *Conceptions of giftedness*. Cambridge: Cambridge University Press.

Galton, F. (1879). Psychometric experiments. *Brain, 2*, 149-162.

Galton, F. (1883). *Inquiries into human faculty and its development*. London: Macmillan.

Galton, F. (1888). *Natural inheritance*. London: Macmillan.

Garb, H. N. (1992). The trained psychologist as expert witness. *Clinical Psychology Review, 12*, 451-467.

Garb, H. N. (1994). Judgment research: Implications for clinical practice and testimony in court. *Applied and Preventive Psychology, 3*, 173-183.

Garb, H. N., Florio, C., & Grove, W. (1998). The validity of the Rorschach and the MMPI: Results from metaanalyses. *Psychological Science, 9*, 402-404.

Gardner, H. (1975). *The shattered mind: The person after brain damage*. New York: Alfred A. Knopf.

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligence*. New York: Basic Books.

Gardner, H. (1986). The waning of intelligence tests. In R. J. Sternberg and D. K. Detterman (eds.), *What is intelligence? Contemporary viewpoints on its nature and definition*. Norwood, NJ: Ablex.

Gardner, H. (1992). Assessment in context: The alternative to standardized testing. In B. R. Gifford & M. C. O'Connor (eds.), *Alternative views of aptitude, achievement, and instruction*. Boston: Klummer.

Gardner, H. (1993). *Multiple intelligences: The theory in practice*. New York: Basic Books.

Gardner, H. (1998). Are there additional intelligences? The case for naturalistic, spiritual, and existential intelligences. In J. Kane (ed.), *Education, information, and transformation*. Englewood Cliffs, NJ: Prentice-Hall.

Gardner, J. (1967). The adjustment of drug addicts as measured by the sentence completion tests. *Journal of Projective Techniques and Personality Assessment, 31*, 28-29.

Gardner, R. A. (1981). Digits forward and digits backward as two separate tests: Normative data on 1567 school children. *Journal of Clinical Child Psychology, 10*, 131-135.

Gardner, W., Lidz, C., Mulvey, E., & Shaw, E. (1996). Clinical versus actuarial predictions of violence in patients with mental illnesse. *Journal of Consulting and Clinical Psychology, 64*, 602-609.

Gass, C., & Brown, M. (1992). Neuropsychological test feedback to patients with brain dysfunction. *Psychological Assessment, 4*, 272-277.

Gazzaniga, M. S. (1970). *The bisected brain*. New York: Appleton-Century-Crofts.

Gazzaniga, M. S., & LeDoux, J. E. (1978). *The integrated mind*. New York: Plenum Press.

Geary, D. C., & Whitworth, R. H. (1988). Is the factor structure of the WISC-R different for Anglo- and Mexican-American children? *Journal of Psychoeducational Assessment, 6*, 253-260.

GED Testing Service (1991). *Examiner's manual: Test of General Educational Development*. Washington, DC: GED Testing Service of the American Council on Education.

Geffen, G., Moar, K. J., O'Hanlon, A. P., Clark, C. R., & Geffen, L. B. (1990). Performance measures of 16-to 86-year-old males and females on the Auditory Verbal Learning Test. *The Clinical Neuropsychologist, 4*, 45-64.

Gelb, S. (1986). Henry H. Goddard and the immigrants, 1910-1917: The studies and their social context. *Journal of the History of the Behavioral Sciences, 22*, 324-332.

Genia, V. (1990). Religious development: A synthesis and reformulation. *Journal of Religion and Health, 29*, 85-99.

Genia, V. (1993a). A psychometric evaluation of the Allport-Ross I/E Scales in a religiously heteroge-

neous sample. *Journal for the Scientific Study of Religion, 32*, 284-290.

Genia, V. (1993b). The Spiritual Experience Index: A measure of spiritual maturity. *Journal of Religion and Health, 30*, 337-347.

Gerard, A. B. (1993). *Manual for Parent-Child Relationship Inventory*. Los Angeles, CA: Western Psychological Services.

Geschwind, N. (1972). Language and the brain. *Scientific American, 226*, 76-83.

Geschwind, N., & Galaburda, A. M. (1987). *Cerebral lateralization: Biological mechanisms, associations, and pathology*. Cambridge, MA: MIT Press.

Gesell, A., Ilg, F. L., and Ames, L. B. (1974). *Infant and child in the culture of today* (rev. ed.). New York: Harper & Row.

Getter, H., & Nowinski, J. (1981). A free response test of interpersonal effectiveness. *Journal of Personality Assessment, 45*, 301-308.

Getz, I. R. (1984). Moral judgment and religion: A review of the literature. *Counseling and Values, 28*, 94-116.

Ghez, C. (1991). The cerebellum. In E. R. Kandel, J. H. Schwartz, & T. M. Jessell (eds.), *Principles of neural science* (3rd ed.). New York: Elsevier.

Ghiselli, E. E. (1966). *The validity of occupational aptitude tests*. New York: Wiley.

Ghiselli, E. E. (1973). The validity of aptitude tests in personnel selection. *Personnel Psychology, 26*, 461-477.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences*. San Francisco: W. H. Freeman.

Gifford, R. (1991). *Applied psychology: Variety and opportunity*. Boston: Allyn & Bacon.

Gilberstadt, H., and Duker, J. (1965). *A handbook for clinical and actuarial MMPI interpretation*. Philadelphia: W. B. Saunders.

Gilliam, J. E. (1994). *Attention-Deficit/Hyperactivity Disorder Test examiner's manual*. Austin, TX: PRO-ED.

Ginsburg, H., and Opper, S. (1988). *Piaget's theory of intellectual development* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Glascoe, F. P. (1991). Developmental screening: Rationale, methods and application. *Infants and Young Children, 4*, 1-10.

Glascoe, F. P., & Byrne, K. E. (1993). The accuracy of three developmental screening tests. *Journal of Early Intervention, 17*, 368-379.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. *American Psychologist, 18*, 519-522.

Goddard, H. H. (1910a). A measuring scale for intelligence. *The Training School, 6*, 146-155.

Goddard, H. H. (1910b). Four hundred feebleminded children classified by the Binet method. *Pedagogical Seminary, 17*, 387-397.

Goddard, H. H. (1911). Two thousand normal children measured by the Binet measuring scale of intelligence. *Pedagogical Seminary, 18*, 232-259.

Goddard, H. H. (1912). *Feeble-mindedness and immigration*. Training School Bulletin, *9*, 91.

Goddard, H. H. (1917). The mental level of a group of immigrants. *Psychological Bulletin, 14*, 68-69.

Goddard, H. H. (1928). Feeblemindedness: A question of definition. *Journal of Psycho-Asthenics, 33*, 219-227.

Goffin, R. D., Rothstein, M., & Johnston, N. (1996). Personality testing and the assessment center: Incremental validity for managerial selection. *Journal of Applied Psychology, 81*, 746-756.

Gold, J., Randolph, C., Carpenter, C., and others (1992). The performance of patients with schizophrenia on the Wechsler Memory Scale-Revised. *Clinical Neuropsychologist, 6*, 367-373.

Goldberg, L. R. (1959). The effectiveness of clinicians' judgments: The diagnosis of organic brain damage from the Bender-Gestalt Test. *Journal of Consulting Psychology, 23*, 25-33.

Goldberg, L. R. (1965). Diagnosticians vs. diagnostic signs: The diagnosis of psychosis vs. neurosis from the MMPI. *Psychological Monographs, 79* (9, Whole No. 602).

Goldberg, L. R. (1981a). Developing a taxonomy of trait-descriptive terms. In D. Fiske (ed.), *New directions for methodology of social and behavioral science: Problems with language imprecision* (no. 9). San Francisco: Jossey-Bass.

Goldberg, L. R. (1981b). Language and individual differences: The search for universals in personality lexicons. In L. Wheeler (ed.), *Review of personality and social psychology*. Beverly Hills, CA: Sage.

Goldberg, L. R. (1990). An alternative "description of personality": The big-five factor structure. *Journal of Personality and Social Psychology, 59*, 1216-1229.

Goldberg, P. (1965). A review of sentence completion methods in personality assessment. In B. I. Murstein (ed.), *Handbook of projective techniques*. New York: Basic Books.

Golden, C. J. (1989). The Luria-Nebraska Neuropsychological Battery. In C. S. Newmark (ed.), *Major psychological assessment instruments* (vol. II). Boston: Allyn & Bacon.

Golden, C. J., Moses, J. A., Graber, B., & Berg, R. A. (1981). Objective clinical rules for interpreting the Luria-Nebraska Neuropsychological Battery: Derivation, effectiveness, and validation. *Journal of Consulting and Clinical Psychology, 49*, 616-618.

Golden, C. J., Purish, A. D., & Hammeke, T. A. (1980). *Luria-Nebraska Neuropsychological Battery: Manual*. Los Angeles: Western Psychological Services.

Golden, C. J., Purish, A. D., & Hammeke, T. A. (1986). *Luria-Nebraska Neuropsychological Battery: Forms I and II*. Los Angeles: Western Psychological Services.

Goldfried, M. R., & Zax, M. (1965). The stimulus value of the TAT. *Journal of Projective Techniques, 29*, 46-57.

Golding, S. L., Roesch, R., and Schreiber, J. (1984). Assessment and conceptualization of competency to stand trial: Preliminary data on the Interdisciplinary Fitness Interview. *Law and Human Behavior, 8*, 321-334.

Goldstein, D., Fogle, E., Wieber, J., & O'Shea, T. (1995). Comparison of the Bayley Scales of Infant Development-Second Edition and the Bayley Scales of Infant Development with premature infants. *Journal of Psychoeducational Assessment, 13*, 391-396.

Goldstein, I. L. (1991). Training in work organizations. In M. D. Dunnette & L. M. Hough (eds.), *Handbook of industrial and organizational psychology* (vol. 2). Palo Alto, CA: Consulting Psychologists Press.

Goldstein, I. L. (1992). *Training* (3rd ed.). Monterey, CA: Brooks/Cole.

Goldstein, K. (1944). The mental changes due to frontal lobe damage. *Journal of Psychology, 17*, 187-208.

Gonzales, E. (1986). *Human figure drawing test*. Austin, TX: PRO-ED.

Goodenough, F. L. (1926). *Measurement of intelligence by drawings*. New York: Harcourt, Brace & World.

Goodenough, F. L. (1949). *Mental testing: Its history, principles, and applications*. New York: Rinehart.

Goodglass, H. (1986). The flexible battery in neuropsychological assessment. In T. Incagnoli, G. Goldstein, & C. J. Golden (eds.), *Clinical applications of neuropsychological test batteries*. New York: Plenum Press.

Goodglass, H., & Kaplan, E. (1983). *The assessment of aphasia and related disorders*. Philadelphia: Lea & Febiger.

Goodman, J. (1990). Infant intelligence: Do we, can we, should we assess it? In C. C. Reynolds, and R. W. Kamphaus (eds.), *Handbook of psychological and educational assessment of children: Intelligence and achievement*. New York: Guilford.

Goodman, J., Malizia, K., Durieux-Smith, A., MacMurray, B., and Bernard, P. (1990). Bayley Developmental Performance at two years of age of neonates at risk for hearing loss. *Developmental Medicine and Child Neurology, 32*, 689-697.

Gordon, M., & Mettelman, B. B. (1988). The assessment of attention: I. Standardization and reliability of a behavior-based measure. *Journal of Clinical Psychology, 44*, 688-690.

Gordon, R., and Peek, L. A. (1989). *The Custody Quotient*. Dallas, TX: Wilmington Institute.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Goslin, D. A. (1963). *The search for ability: Standardized testing in social perspective*. New York: Russell Sage Foundation.

Gothard, S., Viglione, D., Meloy, J. R., & Sherman, M. (1996). Detection of malingering in competency to stand trial evaluations. *Law and Human Behavior, 19*, 493-505.

Gottfredson, G. D. (1990). *Applications and research using Holland's theory of careers: Where we would like to be and suggestions for getting there*. Paper presented at the annual meeting of the American Psychological Association, Boston, MA.

Gottfredson, G. D., & Holland, J. L. (1975). Vocational choices of men and women: A comparison of predictors from the Self-Directed Search. *Journal of Counseling Psychology, 22*, 28-34.

Gottfredson, G. D., & Holland, J. L. (1989). *Dictionary of Holland Occupational Codes* (2nd ed.). Odessa, FL: Psychological Assessment Resources.

Gottfredson, L. S. (1986). Societal consequences of the *g* factor in employment. *Journal of Vocational Behavior, 29*, 379-410.

Gough, H. G. (1971). The assessment of wayward impulse by means of the Personnel Reaction Blank. *Personnel Psychology, 24*, 669-677.

Gough, H. G. (1984). A managerial potential scale for the California Psychological Inventory. *Journal of Applied Psychology, 69*, 233-244.

Gough, H. G. (1987). *California Psychological Inventory manual*. Palo Alto, CA: Consulting Psychologists Press.

Gough, H. G., & Bradley, P. (1992a). Comparing two strategies for developing personality scales. In M. Zeidner & R. Most (eds.), *Psychological testing: An inside view*. Palo Alto, CA: Consulting Psychologists Press.

Gough, H. G., & Bradley, P. (1992b). Delinquent and criminal behavior as assessed by the Revised California Psychological Inventory. *Journal of Clinical Psychology, 48*, 298-307.

Gould, S. J. (1981). *The mismeasure of man*. New York: Norton.

Graham, J. R. (1987). *The MMPI: A practical guide* (2nd ed.). New York: Oxford University Press.

Graham, J. R. (1993). *MMPI-2: Assessing personality and psychopathology*. New York: Oxford.

Granstrom, S. L. (1987). *A comparative study of loneliness, Buberian religiosity and spiritual well-being in cancer patients*. Paper presented at the conference of the National Hospice Organization.

Graybill, D., & Heuvelman, L. (1993). Validity of the Children's Picture-Frustration Study: A social-cognitive perspective. *Journal of Personality Assessment, 60*, 379-389.

Greenough, W. T., Black, J. E., & Wallace, C. S. (1987). Experience and brain development. *Child Development, 58*, 539-559.

Gregory, R. J. (1987). *Adult intellectual assessment*. Boston, MA: Allyn & Bacon.

Gregory, R. J. (1994a). Aptitude tests. In R. J. Sternberg (ed.), *Encyclopedia of human intelligence*. New York: Macmillan.

Gregory, R. J. (1994b). Profile interpretation. In R. J. Sternberg (ed.), *Encyclopedia of human intelligence*. New York: Macmillan.

Gregory, R. J. (1998). Testing in clinical psychology. In S. Cullari (ed.), *Foundations of clinical psychology*. Boston: Allyn & Bacon.

Gregory, R. J. (1999). *Foundations of intellectual assessment: The WAIS-III and other tests in clinical practice*. Boston: Allyn & Bacon.

Gregory, R. J., & Gernert, C. H. (1990). Age trends for fluid and crystallized intelligence in an able subpopulation. Unpublished manuscript.

Gregory, R. J., & Mohan, P. J. (1977). Effects of asymptomatic lead exposure on childhood IQ: A review. *Intelligence, 1*, 381-400.

Gregory, R. J., Alley, P., & Morris, L. (1980). Lefthandedness and spatial reasoning abilities: The deficit hypothesis revisited. *Intelligence, 4*, 151-159.

Gregory, R. J., Lehman, R. E., & Mohan, P. J. (1976). Intelligence scores for children with and without undue lead absorption. In G. Wegner (ed.), *Shoshone lead health project* (pp. 120-150). Boise: Idaho Department of Health and Welfare.

Gresham, F. M. (1993). "What's wrong in this picture?": Response to Motta et al.'s review of human figure drawings. *School Psychology Quarterly, 8*, 182-186.

Gresham, F. M., & Elliott, S. N. (1990). *Social skills rating system manual*. Circle Pines, MN: American Guidance Service.

Griffith, P., Myers, R., Cusick, G., & Tankersley, M. (1997). MMPI-2 profiles of women differing in sexual abuse history and sexual orientation. *Journal of Clinical Psychology, 53*, 791-800.

Gronwall, D. M. A. (1977). Paced auditory serial-addition task: a measure of recovery from concussion. *Perceptual and Motor Skills, 44*, 367-373.

Gronwall, D. M. A., & Sampson, H. (1974). *The psychological effects of concussion*. Auckland, NZ: Auckland University Press/Oxford University Press.

Gronwall, D. M. A., & Wrightson, P. (1974). Delayed recovery of intellectual function after minor head injury. *Lancet, 2*, 605-609.

Gronwall, D. M. A., & Wrightson, P. (1981). Memory and information processing capacity after closed head injury. *Journal of Neurology, Neurosurgery, and Psychiatry, 44*, 889-895.

Groth-Marnat, G. (1990). *Handbook of psychological assessment*. New York: Wiley.

Groth-Marnat, G. (1997). *Handbook of psychological assessment* (2nd ed.). New York: Wiley.

Gudjonsson, G. H. (1995). The Standard Progressive Matrices: methodological problems associated with the administration of the 1992 adult standardisation sample. *Personality and Individual Differences, 18*, 441-442.

Guilford, J. P. (1959). *Personality*. New York: McGraw-Hill.

Guilford, J. P. (1967). *The nature of human intelligence*. New York: McGraw-Hill.

Guilford, J. P. (1985). The Structure-of-Intellect model. In B. B. Wolman (ed), *Handbook of intelligence: Theories, measurements, and applications*. New York: Wiley.

Guilford, J. P., & Fruchter, B. (1978). *Fundamental statistics in psychology and education* (6th ed.). New York: McGraw-Hill.

Guilford, J. P., & Hoepfner, R. (1971). *The analysis of intelligence*. New York: McGraw-Hill.

Guion, R. M. (1965). *Personnel testing*. New York: McGraw-Hill.

Guion, R. M. (1980). On trinitarian doctrines of validity. *Professional Psychology, 11*, 385-398.

Guion, R. M. (1991). Personnel Assessment, selection, and placement. In M. D. Dunnette & L. M. Hough (eds.), *Handbook of industrial and organizational psychology* (vol. 2). Palo Alto, CA: Consulting Psychologists Press.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Gutkin, R. B., & Reynolds, C. R. (1981). Factorial similarity of the WISC-R for white and black children from the standardization sample. *Journal of Educational Psychology, 73*, 227-231.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*, 139-150.

Guttman, L. (1947). The Cornell technique for scale and intensity analysis. *Educational and Psychological Measurement, 7*, 247-280.

Gynther, M. D., & Gynther, R. A. (1976). Personality inventories. In I. B. Weiner (ed.), *Clinical methods in psychology*, New York: Wiley.

Haaland, K. Y., & Delaney, H. D. (1981). Motor deficits after left or right hemisphere damage due to stroke or tumor. *Neuropsychologia, 19*, 17-27.

Hachinski, V. C., Iliff, L., Zilha, E., & others (1975). Cerebral blood flow in dementia. *Archives of Neurology, 32*, 632-637.

Haddad, F. A. (1986). Comparison of the WISC-R, PPVT-R, and PPVT for learning disabled children. *Psychological Reports, 58*, 659-662.

Haier, R. J., Nuechterlein, K., Hazlett, E., and others. (1988). Cortical glucose metabolic rate correlates of abstract reasoning and attention

studied with positron emission tomography. *Intelligence, 12*, 199-217.

Haier, R. J., Siegel, B., Tang, C., and others. (1992). Intelligence and changes in regional cerebral glucose metabolic rate following learning. *Intelligence, 16*, 415-426.

Hain, J. (1964). The Bender-Gestalt Test: A scoring method for identifying brain damage. *Journal of Consulting and Clinical Psychology, 28*, 34-40.

Haladyna, T. M. (1992). Review of the Millon Clinical Multiaxial Inventory-II. *Eleventh mental measurements yearbook*. Lincoln, NE: University of Nebraska Press.

Hall, H. V. (1987). *Violence prediction: Guidelines for the forensic practitioner*. Springfield, IL: Charles C. Thomas.

Halpern. D. F. (1986). *Sex differences in cognitive abilities*. Hillsdale, NJ: Erlbaum.

Hambleton, R. K. (1984). Validating the test scores. In R. A. Berk (ed.), *A guide to criterion-referenced test construction*. Baltimore, MD: The Johns Hopkins University Press.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (ed.), *Educational measurement* (3rd ed.). New York: American Council on Education/ Macmillan.

Hamilton, S. E. (1993). Identifying African American gifted children using a behavioral assessment technique: The Gifted Child Locator. *Journal of Black Psychology, 19*, 63-76.

Hammeke, T. A., Golden, C. J., & Purish, A. D. (1978). A standardized, short and comprehensive neuropsychological test battery based on the Luria neuropsychological evaluation. *International Journal of Neuroscience, 8*, 135-141.

Hammill, D. D. (1990). On defining learning disabilities: An emerging consensus. *Journal of Learning Disabilities, 23*, 74-84.

Hammill, D. D. (1999). *Detroit Tests of Learning Aptitude-4 (DTLA-4)*. Austin, TX: PRO-ED.

Hansen, J. C. (1984). *User's guide for the SVIB-SCII*. Stanford, CA: Stanford University Press.

Hansen, J. C. (1987). Computer-assisted interpretation of the Strong-Campbell Interest Inventory. In J. N. Butcher (ed.), *Computerized psychological assessment*. New York: Basic Books.

Hansen, J. C. (1992). *Strong user's guide, Revised edition*. Palo Alto, CA: Consulting Psychologists Press.

Hansen, J. C., & Campbell, D. P. (1985). *Manual for the Strong Interest Inventory Form T325 of the Strong Vocational Interest Blanks, Fourth Edition*. Stanford, CA: Stanford University Press.

Hanson, G. A. (1991). To catch a thief: The legal and policy implications of honesty testing in the workplace. *Law and Inequality, 9*, 497-531.

Hargrave, G. E. (1985). Using the MMPI and CPI to screen law enforcement applicants: A study of reliability and validity of clinicians' decisions. *Journal of Police Science and Administration, 13*, 221-224.

Hargrave, G., & Hiatt, D. (1989). Use of the California Psychological Inventory in law enforcement officer selection. *Journal of Personality Assessment, 53*, 267-277.

Hargrave, G., Hiatt, D., Ogard, E., & Karr, C. (1994). Comparison of the MMPI and the MMPI-2 for a sample of peace officers. *Psychological Assessment, 6*, 27-32.

Harmon, L. W. (1989). Counseling. In R. L. Linn (ed.), *Educational measurement* (3rd ed.). New York: American Council on Education/Macmillan.

Harmon, L. W., Hansen, J. C., Borgen, F., & Hammer, A. (1994). *Strong Interest Inventory applications and technical guide*. Palo Alto, CA: Consulting Psychologists Press.

Harrell, T. H., Honaker, L., & Parnell, T. (1992). Equivalence of the MMPI-2 with the MMPI in psychiatric patients. *Psychological Assessment, 4*, 460-465.

Harrington, D. M. (1975). Effect of explicit instructions to "be creative" on the psychological meaning of divergent thinking test scores. *Journal of Personality, 43*, 434-454.

Harrington, R. G. (ed.). (1986). *Testing adolescents*. Kansas City, MO: Test Corporation of America.

Harris, D. B. (1963). *Children's drawings as measures of intellectual maturity*. New York: Harcourt, Brace & World.

Harris, M. M., & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self –peer, and peer-supervisor ratings. *Personnel Psychology, 38*, 43-62.

Harrison, D. A., & Hulin, C. L. (1989). Investigations of absenteeism: Using event history models to study the absence-taking process. *Journal of Applied Psychology, 74*, 300-316.

Harrison, D. A., & Shaffer, M. (1994). Comparative examinations of self-reports and perceived absenteeism norms: Wading through Lake Wobegon. *Journal of Applied Psychology, 79*, 240-251.

Harrison, P. L., & Schock, H. H. (1994). Draw-A-Figure test. In R. J. Sternberg (ed.), *Encyclopedia of human intelligence*, New York: Macmillan.

Hart, K. E. (1997). A moratorium on research using the Jenkins Activity Survey for Type A behavior? *Journal of Clinical Psychology, 53*, 905-907.

Harvey, R. J. (1990). *The common-metric questionnaire for the analysis and evaluation of jobs* (field test version 1.12). San Antonio, TX: The Psychological Corporation.

Harvey, R. J. (1991). Job analysis. In M. D. Dunnette & L. M. Hough (eds.), *Handbook of industrial and*

organizational psychology (vol. 2). Palo Alto, CA: Consulting Psychologists Press.

Harvey, R. J., Friedman, L., Hakel, M., & Cornelius, E. (1988). Dimensionality of the Job Element Inventory, a simplified worker-oriented job analysis questionnaire. *Journal of Applied Psychology, 73*, 639-646.

Hathaway, S. R., & McKinley, J. C. (1940). A multiphasic personality schedule (Minnesota): I. Construction of the schedule. *Journal of Psychology, 10*, 249-254.

Hathaway, S. R., & McKinley, J. C. (1942). A Multiphasic Personality Schedule (Minnesota): III. The measurement of symptomatic depression. *Journal of Psychology, 14*, 73-84.

Hathaway, S. R., & McKinley, J. C. (1943). *The Minnesota Multiphasic Personality Inventory* (rev. Ed.). Minneapolis, MN: University of Minnesota Press.

Hawkins, D. B. (1988). Interpersonal behavior traits, spiritual well-being, and their relationships to blood pressure (doctoral dissertation, Western Conservative Baptist Seminary, 1986). *Dissertation Abstracts International, 48*, 3680B.

Hawkins, D. B., & Larson, R. (1984). *The relationship between measures of health and spiritual well-being.* Unpublished manuscript, Western Conservative Baptist Seminary, Portland, OR.

Hawkins, K. A., Faraone, S. V., Pepple, J. R., Seidman, L. J., & Tsuang, M. T. (1990). WAIS-R validation of the Wonderlic Personnel Test as a brief intelligence measure in a psychiatric sample. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 2*, 198-201.

Haynes, S. G., Feinleib, M., & Eaker, E. (1983). Type A behavior and the ten-year incidence of coronary heart disease in the Framingham heart study. In R. H. Rosenman (ed.), *Psychosomatic risk factors and coronary heart disease*. Bern, Switzerland: Huber.

Haynes, S. N. (1998). Principles and practices of behavioral assessment with adults. In A. S. Bellack & M. Hersen (eds.), *Comprehensive clinical psychology* (vol. 4). Amsterdam: Elsevier.

Hayslip, B., & Panek, P. E. (1989). *Adult development and aging*. New York: Harper & Row.

Hayworth, M. (1970). H-T-P: A projective device. In O. K. Buros (ed.), *Personality tests and reviews*. Highland Park, NJ: The Gryphon Press.

Heaton, R. K., Chelune, G., Talley, J., and others (1993). *Wisconsin Card Sorting Test manual: Revised and expanded*. Odessa, FL: Psychological Assessment Resources.

Heaton, R. K., Smith, H. H., Jr., Lehman, R. A. W., & Vogt, A. T. (1978). Prospects for faking believable deficits on neuropsychological testing. *Journal of Consulting and Clinical Psychology, 46*, 892-900.

Hebb, D. O. (1939). Intelligence in man after large removals of cerebral tissue: Report of four left frontal lobe cases. *Journal of General Psychology, 21*, 73-87.

Heilbrun, A. B., Jr. (1972). Edwards Personal Preference Schedule. *The Seventh mental measurements yearbook*. Lincoln, NE: University of Nebraska Press.

Heilbrun, A. B., Jr., & Georges, M. (1990). The measurement of principled morality by the Kohlberg Moral Dilemma Questionnaire. *Journal of Personality Assessment, 55*, 183-194.

Heilbrun, K. (1992). The role of psychological testing in forensic assessment. *Law and Human Behavior, 16*, 257-272.

Heinemann, A. W., Harper, R. G., Friedman, L. C., & Whitney, J. (1985). The relative utility of the Shipley-Hartford Scale: Prediction of WAIS-R IQ. *Journal of Clinical Psychology, 42*, 547-551.

Heinze, M., & Grisso, T. (1996). Review of instruments assessing parenting competencies used in child custody evaluations. *Behavioral Sciences and the Law, 14*, 293-313.

Helms, J. E. (1983). *A practitioners guide to the Edwards Personal Preference Schedule*. Springfield, IL: Charles C. Thomas.

Hendrickson, A. E. (1982). The biological basis of intelligence. Part 1: Theory. In H. J. Eysenck, (ed.), *A model for intelligence*. Berlin: Springer-Verlag.

Hendrickson, D. E. (1982). The biological basis of intelligence Part II: Measurement. In H. J. Eysenck (ed.). *A model for intelligence*. New York: Springer-Verlag.

Herman, D. O. (1988). Blind Learning Aptitude Test. In D. J. Keyser & R. C. Sweetland (eds.), *Test critiques* (vol. 5). Kansas City, MO: Test Corporation of America.

Herrnstein, R. J., & Murray, C. (1994). *The bell curve: Intelligence and class structure in American life*. New York: Free Press.

Hersen, M., & Bellack, A. S. (eds.). (1988). *Dictionary of behavioral assessment techniques*. New York: Pergamon.

Hersen, M., & Bellack, A. S. (eds.). (1998). *Behavioral assessment: A practical handbook* (2nd ed.). Boston: Allyn & Bacon.

Hiatt, D., & Hargrave, G. E. (1988). MMPI profiles of problem peace officers. *Journal of Personality Assessment, 52*, 722-731.

Hill, E. F. (1972). *The Holtzman Inkblot Technique: A handbook for clinical application*. San Francisco: Jossey-Bass.

Hilliard, A. G. (1984). IQ testing as the emperor's new clothes: A critique of Jensen's Bias in Mental Testing. In C. R. Reynolds & R. T. Brown (eds.), *Perspectives on bias in mental testing*. New York: Plenum.

Hilsenroth, M. J., Fowler, J., Padawer, J., & Handler, L. (1997). Narcissism in the Rorschach revisited: Some reflections on empirical data. *Psychological Assessment, 9*, 113-121.

Hiskey, M. S. (1966). *Manual for the Hiskey-Nebraska Test of Learning Aptitude*. Lincoln, NE: Union College Press.

Hodges, J. R. (1994). *Cognitive assessment for clinicians*. Oxford: Oxford University Press.

Hoffart, A., Friis, S., Strand, J., & Olsen, B. (1994). Symptoms and cognitions during situational and hyperventilatory exposure in agoraphobic patients with and without panic. *Journal of Psychopathology and Behavioral Assessment, 16*, 15-32.

Hoffman, F. J., Sheldon, K. L., Minskoff, E. H., and others (1987). Needs of learning disabled adults. *Journal of Learning Disabilities, 20*, 43-52.

Hogan, J., & Hogan, R. (1986). *Manual for the Hogan Personnel Selection System*. Minneapolis, MN: National Computer Systems.

Hogan, J., & Hogan, R. (1989). How to measure employee reliability. *Journal of Applied Psychology, 74*, 273-279.

Hogan, J., & Quigley, A. (1994). Effects of preparing for physical ability tests. *Public Personnel Management, 23*, 85-104.

Hogan, R. T. (1986). *Manual for the Hogan Personality Inventory*. Minneapolis, MN: National Computer Systems.

Hogan, R. T. (1991). Personality and personality measurement. In M. D. Dunnette & L. M. Hough (eds.), *Handbook of industrial and organizational psychology* (vol. 2). Palo Alto, CA: Consulting Psychologists Press.

Hoge, D. R. (1996). Religion in America: The demographics of belief and affiliation. In E. P. Shafranske (ed.), *Religion and the clinical practice of psychology*. Washington, DC: American Psychological Association.

Holland, J. L. (1966). *The psychology of vocational choice*. Waltham, MA: Blaisdell.

Holland, J. L. (1978). *The occupations finder*. Palo Alto, CA: Consulting Psychologists Press.

Holland, J. L. (1985a). *Making vocational choices: A theory of vocational personalities and work environments* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Hollan, J. L. (1985b). *Self-Directed Search: Professional manual—1985 edition*. Odessa, FL: Psychological Assessment Resources, Inc.

Holland, J. L. (1985c). *Vocational Preference Inventory (VPI) manual—1985 edition*. Odessa, FL: Psychological Assessment Resources, Inc.

Holland, J. L. (1987). *1987 manual supplement for the Self-Directed Search*. Odessa, FL: Psychological Assessment Resources.

Holland, J. L. (1990). *Applications and research using Holland's theory of careers: Where are we now?* Paper presented at the annual meeting of the American Psychological Association, Boston, MA.

Holland, J. L., & Gottfredson, G. D. (1990). *An annotated bibliography for Holland's theory of vocational personalities and work environments*. Paper presented at the annual meeting of the American Psychological Association, Boston, MA.

Hollingshead, A., & Redlich, F. (1958). *Social class and mental illness*. New York: Wiley.

Hollon, S. D., & Kendall, P. C. (1980). Cognitive self-statements in depression: Development of an automatic thoughts questionnaire. *Cognitive Therapy and Research, 4*, 383-397.

Holmes, T., & Rahe, R. (1967). The Social Readjustment Rating Scale. *Journal of Psychosomatic Research, 11*, 213-218.

Holtzman, W. (1988). Beyond the Rorschach. *Journal of Personality Assessment, 52*, 578-609.

Holtzman, W. H. (1961). *Guide to administration and scoring: Holtzman Inkblot Technique*. New York: The Psychological Corporation.

Holzinger, K. J., & Swineford, F. (1939). *A study in factor analysis: The stability of a bi-factor solution*. University of Chicago, Supplementary Educational Monographs, No. 48.

Holzinger, K., & Harman, H. (1941). *Factor analysis: A synthesis of factorial methods*. Chicago: University of Chicago Press.

Honts, C. R., and Perry, M. V. (1992). Polygraph admissibility: Changes and challenges. *Law and Human Behavior, 16*, 357-379.

Honzik, M. (1957). Developmental studies of parent-child resemblance in intelligence. *Child Development, 28*, 215-228.

Honzik, M. (1983). Measuring mental abilities in infancy: The value and limitations. In M. Lewis (ed.), *Origins of intelligence: Infancy and early childhood* (2nd ed.). New York: Plenum.

Horn, J. L, (1968). Organization of abilities and the development of intelligence. *Psychological Review, 75*, 242-259.

Horn, J. L. (1982). The theory of fluid and crystallized intelligence in relation to concepts of cognitive psychology and aging in adulthood. In: F. I. M. Craik and S. Trehub (eds.). *Aging and cognitive proceses*. New York: Plenum Press.

Horn, J. L. (1985). Remodeling old models of intelligence. In B. B. Wolman (ed.), *Handbook of intelligence: Theories, measurements, and applications*. New York: Wiley.

Horn, J. L. (1994). Theory of fluid and crystallized intelligence. In R. J. Sternberg (ed.), *Encyclopedia of human intelligence*. New York: Macmillan.

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general

intelligences. *Journal of Educational Psychology, 57*, 253-270.

Horowitz, F. D. (1994). Giftedness, In R. J. Sternberg (ed.), *Encyclopedia of human intelligence* (vol. 1). New York: Macmillan.

Horowitz, I., and Willging, T. (1984). *The psychology of law*. Boston: Little, Brown.

Hough, L. M., Eaton, N., Dunnette, M., Kamp, J., & McCloy, R. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities [Monograph]. *Journal of Applied Psychology, 75*, 581-595.

Houts, A. C., Cook, T. D., & Shadish, W. R., Jr. (1986). The person-situation debate: A critical multiplist perspective. *Journal of Personality, 54*, 52-105.

Howell, R. J., and Richards, L. (1989). Review of the Rogers Criminal Responsibility Assessment Scales. *The Tenth Mental Measurements Yearbook*. Lincoln, NE: University of Nebraska Press.

Huffcutt, A. I., & Roth, P. (1998). Racial group differences in employment interview evaluation. *Journal of Applied Psychology, 83*, 179-189.

Hughes, J. L., & McNamara, W. J. (1959). *Manual for the revised Programmer Aptitude Test*. New York: The Psychological Corporation.

Huhtaniemi, P., Haier, R. J., Fedio, P., & Buchsbaum, M. S. (1983). Neuropsychological characteristics of college males who show attention dysfunction. *Perceptual and Motor Skills, 51*, 707-713.

Humphreys, L. G. (1971). Theory of intelligence. In R. Cancro (ed.), *Intelligence: genetic and environmental influences*. New York: Grune & Stratton.

Hunsberger, B. (1995). Religion and prejudice: The role of religious fundamentalism, quest, and right-wing authoritarianism. *Journal of Social Issues, 51*, 113-129.

Hunter, J. E. (1989). *The Wonderlic Personnel Test as a predictor of training success and job performance*. Northfield, IL: E. F. Wonderlic Personnel Test, Inc.

Hunter, J. E. (1994). General Aptitude Test Battery. In R. J. Sternberg (ed.), *Encyclopedia of human intelligence*. New York: Macmillan.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job perfomance. *Psychological Bulletin, 96*, 72-98.

Hunter, J. E., & Schmidt, F. L. (1976). Critical analysis of the statistical and ethical implications of various definitions of test bias. *Psychological Bulletin, 83*, 1053-1071.

Hutt, M. L. (1977). *The Hutt Adaptation of the Bender-Gestalt Test*. New York: Grune & Stratton.

Hutt, M. L., & Briskin, G. J. (1960). *The clinical use of the revised Bender-Gestalt Test*. New York: Grune & Stratton.

Hynd, G. W., & Willis, W. G. (1985). Neurological foundations of intelligence. In B. B. Wolman (ed.), *Handbook of intelligence: Theories, measurements, and applications*. New York: Wiley.

Iacono, W. G., & Lykken, D. T. (1997). The validity of the lie detector: Two surveys of scientific opinion. *Journal of Applied Psychology, 82*, 426-433.

Igbokwe, N. U. (1989). Reliability of the Jenkins Activity Survey in assessing Type A behavior pattern in middle-aged Nigerians. *Perceptual and Motor Skills, 69*, 1330.

Inwald, R. E. (1988). Five-year follow-up of departmental terminations as predicted by 16 preemployment psychological indicators. *Journal of Applied Psychology, 73*, 703-710.

IPAT (1973). *Measuring intelligence with the Culture Fair Tests: Manual for Scales 2 and 3*, Champaign, IL: Institute for Personality and Ability Testing.

Irwin, P. M. (1992). *Elementary and Secondary Education Act of 1965: FY 1993 Guide To Programs*. Congressional Research Service. Washington, DC: Government Printing Office.

Itard, J. M. G. (1932/1801). The *wild boy of Aveyron*. Trans. By G. & M. Humphrey. New York: Appleton-Century-Crofts.

Ivnik, R., Malec, J., Smith, G., & others (1992). Mayo's Older Americans Normative Studies: Updated AVLT norms for ages 56 to 97. *Clinical Neuropsychologist, 6*, 83-104.

Jackson, D. N. (1970). A sequential system for personality scale development. In C. D. Spielberger (ed.), *Current topics in clinical and community psychology* (vol. 2). Orlando, FL: Academic Press.

Jackson, D. N. (1977). *Jackson Vocational Interest Survey Manual*. Ontario: Research Psychologists Press.

Jackson, D. N. (1984a). *Manual for the Multidimensional Aptitude Battery*. Port Huron, MI: Research Psychologists Press.

Jackson, D. N. (1984b). *Personality Research Form manual*. Port Huron, MI: Research Psychologists Press.

Jackson, D. N. (1986). *Career Directions Inventory manual*. Port Huron, MI: Sigma Assessment Systems.

Jackson, D. N. (1991). *Jackson Vocational Interest Survey Manual* (3rd ed.). Port Huron, MI: Research Psychologists Press.

Jackson, D. N., & Gray, A. (1987). *The Survey of Work Styles*. Port Huron, MI: Research Psychologists Press.

Jackson, D. N., & Messick, S. (1968). Creativity. In P. London & D. Rosenhan (eds.). *Foundations of abnormal psychology*. New York: Holt.

James, W. (1902). *The varieties of religious experience*. New York: Longman.

Jenkins, C. D., Zyzanski, S., & Rosenman, R. (1971). Progress toward validation of a computer-scored

test for the Type A coronary-prone behavior pattern. *Psychosomatic Medicine, 33*, 193-201.

Jenkins, C. D., Zyzanski, S., & Rosenman, R. (1979). *Jenkins Activity Survey: Manual*. New York: The Psychological Corporation.

Jenkins, J. J., & Paterson, D. G. (eds.). (1961). *Studies in individual differences*. New York: Appleton-Century-Crofts.

Jennett, B., & Teasdale, G. (1981). *Management of head injuries*. Philadelphia PA: F. A. Davis.

Jennett, B., Teasdale, G. M., & Knill-Jones, R. P. (1975). Predicting outcome after head injury. *Journal of Royal College of Physicians of London, 9*, 231-237.

Jensen, A. R. (1969). How much can we boost IQ and scholastic achievement? *Harvard Educational Review, 39*, 1-123.

Jensen, A. R. (1977). Cumulative deficit in IQ of blacks in the rural south. *Developmental Psychology, 13*, 184-191.

Jensen, A. R. (1979). *G*: outmoded theory or unconquered frontier? *Creative Science and Technology, 2*, 16-29.

Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.

Jensen, A. R. (1981). Raising the IQ: The Ramey and Haskins Study. *Intelligence, 5*, 29-40.

Jensen, A. R. (1982). The chronometry of intelligence. In R. J. Sternberg (ed.), *Advances in the psychology of human intelligence*. Hillsdale, NJ: Erlbaum.

Jensen, A. R. (1984). Test bias: Concepts and criticisms. In C. R. Reynolds & R. T. Brown (eds.), *Perspectives on bias in mental testing*. New York: Plenum.

Jensen, A. R., & Osborne, R. T. (1979). *Forward and backward digit span interaction with race and IQ: A longitudinal developmental comparison*. Berkeley: University of California. (ERIC Document Reproduction Service No. ED 173 384).

Johnson, J. H., and Williams, T. A. (1975). The use of online computer technology in a mental health admitting system. *American Psychologist, 30*, 388-390.

Johnson, N. E., Saccuzzo, D., & Guertin, T. (1994). The development and validation of a reliable alternate form for Raven's Standard Progressive Matrices. *Assessment, 1*, 315-319.

Johnson, R. C., McClearn, G. E., Yuen, S., Nagoshi, C. T., Ahern, F. M., & Cole, R. E. (1985). Galton's data a century later. *American Psychologist, 40*, 875-892.

Johnson, R. G. (1987). Shipley Institute of Living Scale. In D. J. Keyser & R. C. Sweetland (eds.), *Test critiques compendium*. Kansas City, MO: Test Corporation of America.

Johnson, S. T. (1994). Scholastic Assessment Tests. In R. J. Sternberg (ed.), *Encyclopedia of human intelligence*. New York: Macmillan.

Johnson, W. G., and Mullett, N. (1987). Georgia Court Competency Test-Revised. In M. Hersen and A. S. Bellack (eds.), *Dictionary of behavioral assessment techniques*. Elmsford, NY: Pergamon.

Johnston, D. W. (1986). Behavior therapy. In R. Harre & R. Lamb (eds.), *The dictionary of physiological and clinical psychology*. Cambridge, MA: The MIT Press.

Johnston, W. T., & Bolen, R. M. (1984). A comparison of the factor structures of the WISC-R for Blacks and Whites. *Psychology in the Schools, 21*, 42-44.

Joint Committee on Testing Practices (1988). *Code of fair testing practices in education*. Washington, DC: Author.

Jones, K. L., Smith, D. W., Ulleland, C. N., & Streissguth, A. P. (1973). Patterns of malformation in offspring of chronic alcoholic mothers. *Lancet, 1*, 1267-1271.

Jones, R. (1980). *The big switch: New careers, new lives after 35*. New York: McGraw-Hill.

Joseph, R. (1988). The right cerebral hemisphere: Emotion, music, visual-spatial skills, body-image, dreams, and awareness. *Journal of Clinical Psychology, 44*, 630-763.

Jung, C. G. (1910). The association method. *American Journal of Psychology, 21*, 219-269.

Kaiser, H. F., & Michael, W. B. (1975). Domain validity and generalizability. *Educational and Psychological Measurement, 35*, 31-35.

Kalichman, S. C. (1988). Empirically derived MMPI profile subgroups of incarcerated homicide offenders. *Journal of Clinical Psychology, 44*, 733-738.

Kamphaus, R., Beres, K., Kaufman, A., & Kaufman, N. (1996). The Kaufman Assessment Battery for Children. In C. S. Newmark (ed.), *Major psychological assessment instruments* (2nd ed.). Boston: Allyn & Bacon.

Kamphaus, R. W. (1990). K-ABC theory in historical and current contexts. *Journal of Psychoeducational Assessment, 8*, 356-368.

Kamphaus, R. W. (1993). *Clinical assessment of children's intelligence*. Boston: Allyn & Bacon.

Kandel, E. R. (1991). Perception of motion, depth, and form. In E. R. Kandel, J. H. Schwartz, & T. M. Jessell (eds.), *Principles of neural science* (3rd ed.). New York: Elsevier.

Kandel, E. R., Schwartz, J. H., & Jessell, T. M. (1995). *Essentials of neural science and behavior*. Norwalk, CN: Appleton & Lange.

Kane, R. L. (1991). Standardized and flexible batteries in neuropsychology: An assessment update. *Neuropsychology Review, 2*, 281-339.

Kaplan, C. (1996). Predictive validity of the WPPSI-R: A four year follow-up study. *Psychology in the Schools, 33*, 211-219.

Kaplan, S. L., & Alfonso, V. C. (1997). Confirmatory factor analysis of the Stanford-Binet Intelligence

Scale: Fourth Edition with preschoolers with developmental delays. *Journal of Psychoeducational Assessment, 15*, 226-237.

Karr, S., Carvajal, H., Elser, D., & Bays, K. (1993). Concurrent validity of the WPPSI-R and the McCarthy Scales of Children's Abilities. *Psychological Reports, 72*, 940-942.

Kaufman, A. S. (1975). Factor structure of the McCarthy Scales at five age levels between 2½ and 8 ½ *Educational and Psychological Measurement, 35*, 641-656.

Kaufman, A. S. (1978). Review of Columbia Mental Maturity Scale. *The Eighth Mental Measurements Yearbook*. Lincoln, NE: University of Nebraska Press.

Kaufman, A. S. (1983). Test review: WAIS-R. *Journal of Psychoeducational Assessment, 1*, 309-319.

Kaufman, A. S. (1990). *Assessing adolescent and adult intelligence*. Boston: Allyn & Bacon.

Kaufman, A. S. & Kaufman, N. L. (1983). *K-ABC administration and scoring manual*. Circle Pines, MN: American Guidance Service.

Kaufman, A. S., & Kaufman, N. L. (1985). *Kaufman Test of Educational Achievement: Comprehensive Form Manual*. Circle Pines, MN: American Guidance Service.

Kaufman, A. S., & Kaufman, N. L. (1990). *Kaufman Brief Intelligence Test manual*. Circle Pines, MN: American Guidance Service.

Kaufman, A. S., & Kaufman, N. L. (1997). The Kaufman Adolescent and Adult Intelligence Test. In D. P. Flanagan, J. Genshaft, & P. Harrison (eds.), *Contemporary intellectual assessment: Theories, tests, and issues*. New York: Guilford.

Kaufman, A. S., & Wang, J. (1992). Gender, race, and education differences on the K-BIT at ages 4 to 90. *Journal of Psychoeducational Assessment, 10*, 219-229.

Kaufman, A. S., Kamphaus, R. W., & Kaufman, N. L. (1985). The Kaufman Assessment Battery for Children (K-ABC). In C. S. Newmark (ed.), *Major psychological assessment instruments*. Boston: Allyn & Bacon.

Kaufman, A. S., Kaufman, N. L., & Goldsmith, B. Z. (1984). *Kaufman sequential or simultaneous: Leader's kit*. Circle Pines, MN: American Guidance Service.

Kaufman, A. S., Kaufman, N. L., Kamphaus, R. W., & Naglieri, J. A. (1982). Sequential and simultaneous factors at ages 3-12 ½: Developmental changes in neuropsychological dimensions. *Clinical Neuropsychology, 4*, 74-81.

Kaufman, A. S., McLean, J. E., & Lincoln, A. (1996). The relationship of the Myers-Briggs Type Indicator (MBTI) to IQ level and the fluid and crystallized IQ discrepancy on the Kaufman Adolescent and Adult Intelligence Test (KAIT). *Assessment, 3*, 225-239.

Kaufman, A. S., McLean, J. E., & Reynolds, C. R. (1988). Sex, race, residence, region, and education differences on the 11 WAIS-R subtests. *Journal of Clinical Psychology, 44*, 231-248.

Kausler, D. (1991). *Experimental psychology, cognition, and human aging* (2nd ed.). New York: Springer-Verlag.

Kazdin, A. E. (1988). The token economy: A decade later. In G. Davey & C. Cullen (eds.), *Human operant conditioning and behavior modification*. New York: Wiley.

Kazdin, A. E. (1990). Evaluation of the Automatic Thoughts Questionnaire: Negative cognitive processes and depression among children. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 2*, 73-79.

Keenan, J., & Brassell, E. (1975). *Aphasia Language Performance Scales*. Murfreesboro, TN: Pinnacle Press.

Keenan, P. A., & Lachar, D. (1988). Screening preschoolers with special problems: Use of the Personality Inventory for Children (PIC). *Journal of School Psychology, 26*, 1-11.

Keith, T. Z., and Bolen, L. M. (1980). Factor structure of the McCarthy Scales for children experiencing problems in school. *Psychology in the Schools, 17*, 320-326.

Kelley, T. L. (1928). *Crossroads in the mind of man: A study of differentiable mental abilities*. Stanford, CA: Stanford University Press.

Kelly, E. L., & Fiske, D. W. (1951). *The prediction of performance in clinical psychology*. Ann Arbor: University of Michigan Press.

Kendall, P. C., & Hollon, S. S. (1989). Anxious self-talk: Development of the Anxious Self-Statements Questionnaire (ASSQ). *Cognitive Therapy and Research, 13*, 81-91.

Kennedy, W. A., Van de Riet, V., & White, J. C., Jr. (1963). A normative sample of intelligence and achievement of negro elementary school children in the southeast United States. *Monographs of the Society for Research in Child Development, 28* [No. 90], 68.

Kent, G. H., & Rosanoff, A. J. (1910). A study of association in insanity. *American Journal of Insanity, 67*, 37-96; 317-390.

Kertesz, A. (1979). *Aphasia and associated disorders: Taxonomy, localization, and recovery*. New York: Grune & Stratton.

Keyser, D. J., & Sweetland, R. C. (eds.). (1984-1988). *Test Critiques* (volumes I-VI). Kansas City, MO: Test Corporation of America.

Kifer, E. (1985). Review of ACT Assessment Program. *Ninth Mental Measurements Yearbook*. Lincoln, NE: University of Nebraska Press.

Killian, G. A. (1987). House-Tree-Person technique. In D. J. Keyser & R. C. Sweetland (eds.), *Test critiques compendium*. Kansas City, MO: Test Corporation of America.

Kinnear, P. R., & Gray, C. D. (1997). *SPSS for Windows made simple* (2nd ed.). Trowbridge, UK: Psychology Press.

Kinsbourne, M. (1994). Neuropsychology of attention. In D. W. Zaidel (ed.), *Neuropsychology*. San Diego, CA: Academic Press.

Kirkpatrick, L., & Hood, R. (1990). Intrinsic-Extrinsic Religious Orientation: The boon or bane of contemporary psychology of religion? *Journal for the Scientific Study of Religion, 29*, 442-462.

Kleinmuntz, B., and Szucko, J. J. (1984). A field study of the fallibility of polygraphic lie detection. *Nature, 308*, 449-450.

Klieger, D. M., & Franklin, M. E. (1993). Validity of the fear survey schedule in phobia research: A laboratory test. *Journal of Psychopathology and Behavioral Assessment , 15*, 207-218.

Klimoski, R., & Palmer, S. (1994). The ADA and the hiring process in organizations. In S. M. Bruyere & J. O'Keeffe (eds.), *Implications of the Americans with Disabilities Actor for psychology*. New York: Springer.

Kline, P. (1986). *A handbook of test construction*. New York: Metheun.

Kline, P. (1993). *The handbook of psychology testing*. London: Routledge.

Kline, R. B. (1989). Is the Fourth Edition Stanford-Binet a four-factor test? Confirmatory factor analysis of alternative models for ages 2 through 23. *Journal of Psychoeducational Assessment, 7*, 4-13.

Kline, R. B., & Lachar, D. (1992). Evaluation of age, sex, and race bias in the Personality Inventory for Children (PIC). *Psychological Assessment, 4*, 333-339.

Kline, R. B., Guilmette, S., Snyder, J., & Castellanos, M. (1992). Relative cognitive complexity of the Kaufman Assessment Battery for Children (K-ABC) and the WISC-R. *Journal of Psychoeducational Assessment, 10*, 141-152.

Kline, R. B., Lachar, D., & Boersma, D. (1993). Identification of special education needs with the personality inventory for children (PIC): A hierarchical classification model. *Psychological Assessment, 5*, 307-316.

Kline, R. B., Lachar, D., & Gdowski, C. L. (1992). Clinical validity of a Personality Inventory for Children (PIC) profile typology. *Journal of Personality Assessment, 58*, 691-705.

Kline, R. B., Lachar, D., Gruber, C., & Boersma, D. (1994). Identification of special education needs with the Personality Inventory for Children (PIC): A profile-matching strategy. *Assessment, 1*, 301-314.

Klingler, D. E., Miller, D., Johnson, J., and Williams, T. (1977). Process evaluation of an on-line computer-assisted unit for intake assessment of mental health patients. *Behavior Research Methods and Instrumentation, 9*, 110-116.

Klopfer, W. G., & Taulbee, E. S. (1976). Thematic Apperception Test. *Annual Review of Psychology, 27*, 543-567.

Klove, H. (1963). Clinical neuropsychology. In: F. M. Forster (ed.). *The Medical Clinics of North America* New York. Saunders.

Kluckhohn, C. (1951). Values and value orientation in the theory of action. In T. Parsons and E. A. Shils (eds.), *Toward a general theory of action*. Cambridge, MA: Harvard University Press.

Knight, B. C., Baker, E. H., and Minder, C. C. (1990). Concurrent validity of the Stanford-Binet: Fourth Edition and Kaufman Assessment Battery for Children with learning-disabled students. *Psychology in the Schools, 27*, 116-125.

Knight, R. G. (1992). *The neuropsychology of degenerative brain diseases*. Hillsdale, NJ: Erlbaum.

Knobloch, H., Stevens, F., and Malone, A. (1987). *Manual of developmental diagnosis: The administration and interpretation of the Revised Gesell and Amatruda Developmental and Neurologic Examination*. Houston, TX: Developmental Evaluation Materials, Inc.

Knox, H. A. (1914). A scale, based on the work at Ellis Island, for estimating mental defect. *Journal of the American Medical Association, 62*, 741-747.

Knoch, W. R. (1984). Culture Fair Intelligence Test. In D. J. Keyser & R. C. Sweetland (eds.), *Test Critiques* (vol. I). Kansas City, MO: Test Corporation of America.

Kohlberg, L, (1958). *The development of modes of moral thinking and choice in the years ten to sixteen.* Unpublished doctoral dissertation, University of Chicago.

Kohlberg, L. (1981). *Essays on moral development: Vol. 1. The philosophy of moral development*. San Francisco: Harper & Row.

Kohlberg, L. (1984). *Essays on moral development: Vol. 2. The psychology of moral development*. San Francisco: Harper & Row.

Kohlberg, L., & Elfenbein, D. (1975). The development of moral judgments concerning capital punishment. *American Journal of Orthopsychiatry, 45*, 614-639.

Kohlberg, L., & Kramer, R. (1969). Continuities and discontinuities in children and adult moral development. *Human Development, 12*, 225-252.

Kohs, S. C. (1920). The block-design tests. *Journal of Experimental Psychology, 3*, 357-376.

Kolb, B., & Milner, B. (1981). Performance of complex arm and facial movements after focal brain lesions. *Neuropsychologia, 19*, 491-503.

Kolb, B., & Whisahw, I. Q. (1985). *Fundamentals of human neuropsychology* (2nd ed.). New York: Freeman.

Kolb, B., & Whishaw, I. Q. (1990). *Fundamentals of human neuropsychology* (3rd ed.). New York: Freeman.

Kolb, B., Milner, B., & Taylor, L. (1983). Perception of faces by patients with localized cortical excisions. *Canadian Journal of Psychology, 37*, 8-18.

Koltai, D., Bowler, R., & Shore, M. (1996). The Rivermead Behavioral Memory Test and Wechsler Memory Scale-Revised: Relationship to everyday memory impairment. *Assessment, 3*, 443-448.

Koluchova, J. (1972). Severe deprivation in twins: A case study. *Journal of Child Psychology and Psychiatry, 13*, 107-114.

Konold, T. R., Kush, J., & Canivez, G. (1997). Factor replication of the WISC-III in three independent samples of children receiving special education. *Journal of Psychoeducational Assessment, 15*, 123-137.

Koss, E. (1994). Neuropsychology of aging and dementia. In D. W. Zaidel (ed.), *Neuropsychology* (2nd ed.). San Diego, CA: Academic Press.

Kramer, J. J. (1988). Computer-based test interpretation in psychoeducational assessment: An initial appraisal. *Journal of School Psychology, 26*, 143-153.

Krieshok, T. S., & Harrington, R. G. (1985). A review of the Multidimensional Aptitude Battery. *Journal of Counseling and Development, 64*, 87-89.

Krugman, M. (1970). H-T-P: House, Tree, and Person. In O. K. Buros (ed.), *Personality tests and reviews*. Highland Park, NJ: The Gryphon Press.

Krumboltz, J. D. (1991). *Manual for the Career Beliefs Inventory*. Palo Alto, CA: Consulting Psychologists Press.

Kuder, G. F. (1934). *Kuder preference record*. Chicago: Science Research Associates.

Kuder, G. F. (1966). The Occupational Interest Survey. *Personnel and Guidance Journal, 45*, 72-77.

Kuder, G. F. (1975). *General Interest Survey (Form E)— Manual*. Chicago: Science Research Associates.

Kuder, G. F., & Diamond, E. E. (1979). *Kuder Occupational Interest Survey: General manual*. Chicago: Science Research Associates.

Kuder, G. F., & Richardson, M. W. (1937). The theory of estimation of test reliability. *Psychometrika, 2*, 151-160.

Kuncel, N., Campbell, J., & Ones, D. (1998). Validity of the Graduate Record Examination: Estimated or tacitly known? *American Psychologist, 53*, 567-568.

Kupfermann, I. (1991a). Hypothalamus and limbic system: Peptidergic neurons, homeostasis, and emotional behavior. In E. R. Kandel, J. H. Schwartz, & T. M. Jessell (eds.), *Principles of neural science* (3rd ed.). New York: Elsevier.

Kupfermann, I. (1991b). Localization of higher cognitive and affective functions: The association cortices. In E. R. Kander, J. H. Schwartz, & T. M. Jessel (eds.), *Principles of neural science* (3rd ed.). New York: Elsevier.

Kurtines, W., & Greif, E. B. (1974). The development of moral thought: Review and evaluation of Kohlberg's approach *Psychological Bulletin, 81*, 453-470.

La Rue, A. (1992). *Aging and neuropsychological assessment*. New York: Plenum.

Lastsch, L., & Choca, J. (1994). Cluster-branching methodology for adaptive testing and the development of the Adaptive Category Tests. *Psychological Assessment, 6*, 345-351.

Laboratory for Community Psychiatry (1974). *Competency to stand trial and mental illness*. Northvale, NJ: Jason Aronson.

Lachar, D. (1974). *The MMPI: Clinical assessment and automated interpretation*. Los Angeles: Western Psychological Services.

Lachar, D. (1982). *Personality Inventory for Children: Revised format manual supplement*. Los Angeles: Western Psychological Services.

Lachar, D. (1987). Automated assessment of child and adolescent personality. In J. N. Butcher (ed.), *Computerized psychological assessment: A practitioner's guide*. New York: Basic Books.

Lachar, D., & Gdowski, C. L. (1979). *Actuarial assessment of child and adolescent personality: An interpretive guide for the Personality Inventory for Children profile*. Los Angeles: Western Psychological Services.

Lachar, D., Kline, R. B., & Gdowski, C. L. (1987). Respondent psychopathology and interpretive accuracy of the Personality Inventory for Children: The evaluation of a "most reasonable" assumption. *Journal of Personality Assessment, 51*, 165-177.

Lacks, P. (1999). *Bender-Gestalt screening for brain dysfunction* (2nd ed.). New York: Wiley.

Lacks, P., & Newport, K. (1980). A comparison of scoring systems and level of scorer experience on the Bender-Gestalt Test. *Journal of Personality Assessment, 44*, 351-357.

Lah, M. I. (1989). New validity, normative, and scoring data for the Rotter Incomplete Sentences Blank. *Journal of Personality Assessment, 53*, 607-620.

Lah, M. I., & Rotter, J. B. (1981). Changing college student norms on the Rotter Incomplete Sentences Blank. *Journal of Consulting and Clinical Psychology, 49*, 985.

Lai, T. C. (1970). *A scholar in imperial China*. Hong Kong: Kelly & Walsh.

Lam, T. C. M: (1993). Testability: A critical issue in testing language minority students with standardized achievement tests. *Measurement and Evaluation in Counseling and Development, 26*, 179-191.

Lambert, N., Hartsough, C., & Sandoval, J. (1990). *Manual for the Children's Attention and Adjustment*

Survey. Circle Pines, MN: American Guidance Service.

Landy, F. J. (1985). *The psychology of work behavior* (3rd ed.). Homewood, IL: Dorsey Press.

Landy, F. J., & Farr, J. L. (1983). *The measurement of work performance: Methods, theory and applications*. New York: Academic Press.

Landy, F. J., Shankster, L. J., & Kohler, S. S. (1994). Personnel selection and placement. *Annual Review of Psychology, 45*, 261-296.

Lane, S. (1992). Review of the Iowa Tests of Basic Skills. *Eleventh Mental Measurements Yearbook*. Lincoln, NE: University of Nebraska Press.

Lang, W. (1995). Review of Gesell Child Developmental Age Scale. *Twelfh Mental Measurements Yearbook*. Lincoln, NE: University of Nebraska Press.

Lang, W. S. (1992). Review of TEMAS (Tell-Me-A-Story). *The Eleventh Mental Measurements Yearbook*. Lincoln, NE: University of Nebraska Press.

Lansdell, H., & Donnelly, E. F. (1977). Factor analysis of the Wechsler Adult Intelligence Scale Subtests and the Halstead-Reitan Category and Tapping Tests. *Journal of Consulting and Clinical Psychology, 45*, 412-416.

Lanyon, R. I. (1984). Personality assessment. *Annual Review of Psychology, 35*, 667-701.

Larson, G. E. (1994). Armed Services Vocational Aptitude Battery. In R. J. Sternberg (ed.), *Encyclopedia of human intelligence*. New York: Macmillan.

Larson, G. E., & Wolfe, J. (1995). Validity results for *g* from and expanded test base. *Intelligence, 20*, 15-25.

Lassiter, K., & Bardos, A. (1995). The relationship between young children's academic achievement and measures of intelligence. *Psychology in the Schools, 32*, 170-177.

Latham, G. P., & Skarlicki, D. (1995). Criterion-related validity of the situational and patterned behavior description interviews with organizational citizenship behavior. *Human Performance, 8*, 67-80.

Laurent, J., Swerdlik, M., and Ryburn, M. (1992). Review of validity research on the Stanford-Binet Intelligence Scale: Fourth Edition. *Psychological Assessment, 4*, 102-112.

Lavin, C. (1996). The WISC-III and the Stanford-Binet: Fourth Edition: A preliminary study of validity. *Psychological Reports, 78*, 491-496.

LaVoie, A. L. (1987). The Blacky Pictures, In D. J. Keyser & R. C. Sweetland (eds.), *Test critiques compendium*. Kansas City, MO: Test Corporation of America.

Lawsche, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*, 563-575.

Layton, W. (1992). Review of Minnesota Importance Questionnaire. *Eleventh Mental Measurements Yearbook*. Lincoln, NE: University of Nebraska Press.

Ledbetter, M., Smith, L., Vosler-Hunter, W., & Fischer, J. (1991). An evaluation of the research and clinical usefulness of the Spiritual Well-Being Scale. *Journal of Psychology and Theology, 19*, 49-55.

Lee, M. S., Wallbrown, F., & Blaha, J. (1990). Note on the construct validity of the Multidimensional Aptitude Battery. *Psychological Reports, 67*, 1219-1222.

Lefcourt, H. M. (1982). *Locus of control: Current trends in theory and research* (2nd ed.). Hillsdale, NJ: Erlbaum.

Lefebvre, M. F. (1981). Cognitive distortion and cognitive errors in depressed psychiatric and low back pain patients. *Journal of Consulting and Clinical Psychology, 49*, 517-525.

Lehman, R. E. (1978). Symptom contamination of the Schedule of Recent Events. *Journal of Consulting and Clinical Psychology, 46*, 1564-1565.

Leiter, R. G. (1948). *Leiter International Performance Scale*. Chicago: Stoelting Co.

Leiter, R. G. (1979). *Leiter International Performance Scale: Instruction manual*. Chicago: Stoelting Co.

Leli, D. A., and Filskov, S. B. (1984). Clinical detection of intellectual deterioration associated with brain damage. *Journal of Clinical Psychology, 40*, 1435-1441.

Lerner, J. (1993). *Learning disabilities: Theories, diagnosis & teaching strategies*. Boston, MA: Houghton Mifflin.

Levine, S., Elzey, F. F., Thormahlen, P., & Cain, L. F. (1976). *Manual for the T. M. R. School Competency Scales*. Palo Alto, CA: Consulting Psychologists Press.

Levinson, E. M. (1990). Vocational assessment involvement and use of the Self-Directed Search by school psychologists. *Psychology in the Schools, 27*, 217-228.

Lewinsohn, P. M. (1965). Psychological correlates of overall quality of figure drawings. *Journal of Consulting Psychology, 29*, 504-512.

Lewinsohn, P. M., Munoz, R. F., Youngren, M. A., & Zeiss, A. M. (1986). *Control your depression: Reducing depression through learning self-control techniques, relaxation training, pleasant activities, social skills, constructed thinking, planning ahead, and more* (rev. ed.). New York: Prentice-Hall.

Lewis, M., and Brooks-Gunn, J. (1981). Visual attention at three months as a predictor of cognitive functioning at two years of age. *Intelligence, 5*, 131-140.

Lewis, M., and Sullivan, M. W. (1985). Infant intelligence and its assessment. In B. B. Wolman

(ed.), *Handbook of intelligence: Theories, measurements, and applications*. New York: Wiley.

Lezak, M. (1982). The problem of assessing executive functions. *International Journal of Psychology, 17*, 281-297.

Lezak, M. (1983). *Neuropsychological assessment* (2nd ed.). New York: Oxford University Press.

Lezak, M. (1995). *Neuropsychological assessment* (3rd ed.). New York: Oxford University Press.

Lezak, M. D., & O'Brien, K. P. (1990). Chronic emotional, social, and physical changes after traumatic brain injury. In E. D. Bigler (ed.), *Traumatic brain injury: Mechanisms of damage, assessment, intervention, and outcome*. Austin, TX: PRO-ED.

Lichtenberg, P., Manning, Vangel, S., & Ross. T. (1995). Normative and ecological validity data in older urban medical patients: A program of neuropsychological research. *Advances in Medical Psychotherapy, 8*, 121-136.

Lichtenstein, R. (1990). Psychometric characteristics and appropriate use of the Gesell School Readiness Screening Test. *Early Childhood Research Quarterly, 5*, 359-378.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psycology*, No. 140.

Lindal, E., & Stefansson, J. (1993). Mini-Mental State Examination scores: Gender and lifetime psychiatric disorders. *Psychological Reports, 72*, 631-641.

Lindenberger, U., & Baltes, P. (1994). Aging and intelligence. In R. J. Sternberg (ed.), *Encyclopedia of human intelligence*. New York: Macmillan.

Lindvall, C. M. (1967). *Measuring pupil achievement and aptitude*. New York: Harcourt, Brace & World.

Lindzey, G. (1959). On the classification of projective techniques. *Psychological Bulletin, 56*, 158-168.

Linn, R. L. (1989). Review of the Iowa Tests of Basic Skills. *Tenth Mental Measurements Yearbook*. Lincoln, NE: University of Nebraska Press.

Lipsitt, P. D. (1970). *Competency Screening Test*. Boston: Competency to Stand Trial and Mental Illness Project.

Lipsitt, P. D., Lelos, D., and McGarry, A. L. (1971). Competency for trial: A screening instrument. *American Journal of Psychiatry, 128*, 137-141.

Lipsitz, J. D., Dworkin, R., & Erlenmeyer-Kimling, L. (1993). Wechsler Comprehension and Picture Arrangement subtests and social adjustment. *Psychological Assessment, 5*, 430-437.

Lishman, W. A. (1978). *Organic psychiatry: The psychological consequences of cerebral disorder*. Oxford: Blackwell Scientific Publications.

Loevinger, J. (1976). *Ego development*. San Francisco: Jossey-Bass.

Loevinger, J. (1979). Construct validity of the sentence completion test of ego development. *Applied Psychological Measurement, 3*, 281-311.

Loevinger, J. (1984). On the self and predicting behavior. In R. A. Zucker, J. Aronoff, & A. I. Rabin (eds.), *Personality and the prediction of behavior*. New York: Academic Press.

Lofquist, L. H., & Dawis, R. V. (1991). *Essentials of person-environment correspondence counseling*. Minneapolis, MN: University of Minnesota Press.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Menlo Park, CA: Addison-Wesley.

Lorr, M., Youniss, R., and Stefic, E. (1991). An inventory of social skills. *Journal of Personality Assessment, 57*, 506-520.

Louttit, C. M., & Browne, C. G. (1947). Psychometric instruments in psychological clinics. *Journal of Consulting Psychology, 11*, 49-54.

Lowman, R. L. (1991). *The clinical practice of career assessment: Interests, abilities, and personality*. Washington, DC: American Psychological Association.

Lowman, R. L., & Leeman G. (1988). The dimensionality of social intelligence: Social abilities, interests, and needs. *Journal of Psychology, 122*, 279-290.

Lubin, B., Wallis, R., & Paine, C. (1971). Patterns of psychological test use in the United States: 1935-1969. *Professional Psychology, 2*, 70-74.

Lubinski, D., Benbow, C., & Ryan, J. (1995). Stability of vocational interests among the intellectually gifted from adolescence to adulthood: A 15-year longitudinal study. *Journal of Applied Psychology, 80*, 196-200.

Lubinski, D., Schmidt, D., Benbow, C. (1996). A 20-year stability analysis of the Study of Values for intellectually gifted individuals from adolescence to adulthood. *Journal of Applied Psychology, 81*, 443-451.

Lukens, J., & Hurrell, R. (1996). A comparison of the Stanford-Binet: IV and the WISC-III with mildly retarded children. *Psychological in the Schools, 33*, 24-27.

Lukin, M. E., Dowd, E. T., Plake, B., and Kraft, R. (1985). Comparing computerized vs. traditional psychological assessment. *Computers in Human Behavior, 1*, 49-58.

Lunz, M., & Bergstrom, B. (1994). Computer adaptive testing: A national pilot study. In M. Wilson (ed.), *Objective measurement: Theory into practice* (vol. 2). Norwood, NJ: Ablex.

Lunz, M., Bergstrom, B., & Wirght, B. (1994). Reliability of alternate computer-adaptive tests. In M. Wilson (ed.), *Objective measurement: Theory into practive* (vol. 2). Norwood, NJ: Ablex.

Luria, A. R. (1966). *Higher cortical functions in man*. New York: Basic Books.

Luria, A. R. (1973). *The working brain*. New York: Basic Books.

Lykken, D. T. (1981). *A tremor in the blood: Uses and abuses of the lie detector.* New York: McGraw-Hill.

Lykken, D. T. (1987, Spring). The validity of tests: Caveat emptor. *Jurimetrics Journal*, 263-270.

Lynn, R. (1987). Japan: Land of the rising IQ. A reply to Flynn. *Bulletin of the British Psychological Society, 40*, 464-468.

Lynn, R. (1993). Nutrition and intelligence. In P. A. Vernon (ed.), *Biological approaches to the study of human intelligence*, Norwood, NJ: Ablex.

Lyon, G. R. (ed.). (1994). *Frames of reference for the assessment of learning disabilities: New views on measurement issues.* Baltimore, MD: Brookes Publishing.

MacAndrew, C. (1965). The differentiation of male alcoholic out-patients from nonalcoholic psychiatric patients by means of the MMPI. *Quarterly Journal of Studies on Alcohol, 26*, 238-246.

Maccoby, E. E., & Jacklin, C. N. (1974). *The psychology of sex differences.* Stanford, CA: Stanford University Press.

Machover, K. (1949). *Personality projection in the drawing of the human figure.* Springfield, IL: Charles C. Thomas.

Machover, K. (1951). Drawing of the human figure: A method of personality investigation. In H. Anderson & G. Anderson (eds.), *An introduction to projective techniques.* New York: Prentice-Hall.

MacLennan, R. N. (1992). *PRF annotated bibliography.* Port Huron, MI: Sigma Assessment Systems.

MacPhillamy, D. J., & Lewinsohn, P. M. (1982). The Pleasant Events Schedule: Studies on reliability, validity, and scale intercorrelation. *Journal of Consulting and Clinical Psychology, 50*, 363-380.

Maddi, S. R. (1989). *Personality theories: A comparative analysis* (5th ed.). Chicago: Dorsey Press.

Mahoney, M., & Arnkoff, D. (1978). Cognitive and self-control therapies. In S. Garfield & A. Bergin (eds.), *Handbook of psychotherapy and behavior change: An empirical analysis.* New York: Wiley.

Maker, C. J. (1996). Identification of gifted minority students: A national problem, needed changes and a promising solution. *Gifted Child Quarterly, 40*, 41-50.

Malcolm, K. K. (1998). Developmental Assessment: Evaluation of Infants and Preschoolers. In H. B. Vance (ed.), *Psychological asessment of children: Best practices for schools and clinical settings* (2nd ed.). New York: Wiley.

Malgady, R. G., Constantino, G., & Rogler, L. H. (1984). Development of a Thematic Apperception Tests (TEMAS) for urban Hispanic children. *Journal of Consulting and Clinical Psychology, 52*, 986-996.

Maloney, M. P., & Ward, M. P. (1979). *Mental retardation and modern society.* New York: Oxford University Press.

Manning, W. H., & Jackson, R. (1984). College entrance examinations: Objective selection or gatekeeping for the economically privileged. In C. R. Reynolds & R. T. Brown (eds.), *Perspectives on bias in mental testing.* New York: Plenum.

Mardell-Czudnowski, C., & Goldenberg, D. S. (1990). *DIAL-R: Developmental Indicators for the Assessment of Learning-Revised.* Circle Pines, MN: American Guidance Service.

Marks, P. A., & Seeman, W. (1963). *The actuarial description of abnormal personality.* Baltimore, MD: Williams & Wilkins.

Markwardt, F. C. (1989). *Manual for the Peabody Individual Achievement Test-Revised.* Circle Pines, MN: American Guidance Service.

Martell, D. A. (1992). Forensic neuropsychology and the criminal law. *Law and Human Behavior, 16*, 313-336.

Martens, B. K. (1992). Review of Conners' Rating Scales. *Eleventh Mental Measurements Yearbook.* Lincoln, NE: University of Nebraska Press.

Martin, J. C. (1994). Birth defects. In R. J. Sternberg (ed.), *Encyclopedia of human intelligence.* New York: Macmillan.

Martuza, V. R. (1977). *Applying norm-referenced and criterion-referenced measurement in education.* Boston: Allyn & Bacon.

Mason, C. F., Lemmon, D., Wayne, K., & Schmidt, R. (1991). Shipley Institute of Living Scale: Formulas for Abstraction Quotients from a normative sample of 580. Sex and Socioeconomic status considered ad additional moderating variables. *Psychological Assessment, 3*, 412-417.

Massoth, N. A. (1985). The McCarthy Scales of Children's Abilities as a predictor of achievement: A five-year follow-up. *Psychology in the Schools, 22*, 10-13.

Massoth, N. A., and Levenson, R. L. (1982). The McCarthy Scales of Children's Abilities as a predictor of reading readiness and reading achievement. *Psychology in the Schools, 19*, 293-296.

Matarazzo, J. D. (1972). *Wechsler's measurement and appraisal of adult intelligence* (5th ed.). Baltimore: Williams & Wilkins.

Matarazzo, J. D. (1983). Computerized psychological testing. *Science, 221*, 323.

Matarazzo, J. D. (1986). Computerized clinical psychological test interpretations: Unvalidated plus all mean and no sigma. *American Psychologist, 41*, 14-24.

Matarazzo, J. D. (1990). Psychological assessment versus psychological testing: Validation from Binet to the school, clinic, and courtroom. *American Psychologist, 45*, 999-1017.

Matarazzo, J. D. (1992). Psychological testing and assessment in the 21 st century. *American Psychologist, 47*, 1007-1018.

Matarazzo, J. D., & Herman, D. O. (1984). Relationship of Education and IQ in the WAIS-R standardization sample. *Journal of Consulting and Clinical Psychology, 52*, 631-634.

Matey, C. (1984). Leiter International Performance Scale. In D. J. Keyser & R. C. Sweetland (eds.), *Test critiques* (vol. 1). Kansas City, MO: Test Corporation of America.

Mathews, K. A., & Haynes, S. G. (1986). Type A behavior pattern and coronary risk: Update and critical evaluation. *American Journal of Epidemiology, 123*, 923-960.

Maurer, S. D., & Fay, C. (1988). Effect of situational interviews, conventional structured interviews, and training on interview rating agreement: an experimental analysis. *Personnel Psychology, 41*, 329-344.

Maxey, J. (1994). American College Test. In R. J. Sternberg (ed.), *Encyclopedia of human intelligence*. New York: Macmillan.

Maxwell, J. K., & Wise, F. (1984). PPVT IQ validity in adults: A measure of vocabulary, not of intelligence. *Journal of Clinical Psychology, 40*, 1048-1053.

Mayeux, R., & Kandel, E. R. (1991). Disorders of language: The aphasias. In E. R. Kandel, J. H. Schwartz, & T. M. Jessel (eds.), *Principles of neural science* (3rd ed.). New York: Elsevier.

Mayfield, E. C. (1964). The selection interview: A reevaluation of published research. *Personnel Psychology, 17*, 239-260.

McAllister, L. W. (1986). *A practical guide to CPI interpretation*. Palo Alto, CA: Consulting Psychologists Press.

McCall, R. B. (1976). Toward an epigenetic conception of mental development in the first three years of life. In M. Lewis (ed.), *Origins of intelligence: Infancy and early childhood*. New York: Plenum.

McCall, R. B. (1979). The development of intellectual functioning in infancy and the prediction of later IQ. In J. D. Osofsky (ed.), *Handbook of infant development*. New York: Wiley.

McCall, R. B. (1983). A conceptual approach to early mental development. In M. Lewis (ed.), *Origins of intelligence* (2nd ed.). New York: Plenum.

McCall, W. A. (1939). *Measurement*. New York: Macmillan.

McCallum, R. S. (1990). Determining the factor structure of the Stanford-Binet: Fourth Edition—the right choice. *Journal of Psychoeducational Assessment, 8*, 436-442.

McCallum, R. S., & Bracken, B. (1997). The Universal Nonverbal Intelligence Test. In D. P. Flanagan, J. Genshaft, & P. Harrison (eds.), *Contemporary intellectual assessment: Theories, tests, and issues*. New York: Guilford.

McCallum, R. S., Karnes, F. A., and Oehler-Stinnett, J. (1985). Construct validity of the K-ABC for gifted children. *Psychology in the Schools, 22*, 254-259.

McCann, J. T., & Suess, J. F. (1988). Clinical applications of the MCMI: The 1-2-3-8 Codetype. *Journal of Clinical Psychology, 44*, 181-185.

McCarney, S. B. (1989). *The attention Deficit Disorders Evaluation Scale: School version technical manual*. Columbia, MO: Hawthorne Educational Services.

McCarthy, D. A. (1972). *Manual for the McCarthy Scales of Children's Abilities*. San Antonio, TX: The Psychological Corporation.

McCormick, E. J., Mecham, R., & Jeanneret, P. (1972). A study of job characteristics and job dimensions as based on the Position Analysis Questionnaire (PAQ). *Journal of Applied Psychology, 56*, 347-368.

McCrae, R. R. (1985). Review of the Defining Issues Test. *Ninth mental measurements yearbook*. Lincoln, NE: University of Nebraska Press.

McCrae, R. R., & Costa, P. T., Jr. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychlogy, 2*, 81-90.

McCrowell, K. L., and Nagle, R. (1994). Comparability of the WPPSI-R and the SB:IV among preschool children. *Journal of Psychoeducational Assessment, 12*, 126-134.

McDermott, P. A., Fantuzzo, J. W., & Glutting, J. J. (1990). Just say no to subtest analysis: A critique on Wechsler theory and practice. *Journal of Psychoeducational Assessment, 8*, 290-302.

McDonald, A., Nussbaum, D. And Bagby, R. (1992). Reliability, validity, and utility of the Fitness Interview Test. *Canadian Journal of Psychiatry, 36*, 480-484.

McGlynn, F. D., & Rose, M. P. (1998). Assessment of anxiety and fear. In A. S. Bellack & M. Hersen (eds.), *Behavioral assessment: A practical handbook* (4th ed.). Boston: Allyn & Bacon.

McGrew, M. W., & Teglasi, H. (1990). Formal characteristics of Thematic Apperception Test stories as indices of emotional disturbance in children. *Journal of Personality Assessment, 54*, 639-655.

McGue, M., Bouchard, T., Iacono, W., & Lykken, D. (1993). Behavior genetics of cognitive ability: A lifespan perspective. In R. Plomin & G. McClearn (eds.), *Nature, nurture, and psycholology*. Washington, DC: American Psychological Association.

McGurk, F. C. J. (1953a). On white and Negro test performance and socio-economic factors. *Journal of Abnormal and Social Psychology, 48*, 448-450.

McGurk, F. C. J. (1953b). Socioeconomic status and culturally-weighted test scores of Negro subjects. *Journal of Applied Psychology, 37*, 276-277.

McGurk, F. C. J. (1967). The culture hypothesis and psychological tests. In R. E. Kuttner (ed.), *Race and modern science*. New York: Social Science Press.

McGurk, F. C. J. (1975). Race differences—twenty years later. *Homo, 26*, 219-239.

McKeachie, W. J. (1984). Does anxiety disrupt information processing or does poor information processing lead to anxiety? *International Review of Applied Psychology, 33*, 187-203.

McKey, R. H. And others (1985). *The impact of Head Start on children, families and communities*. Washington, DC: U.S. Government Printing Office.

McKinley, J. C., & Hathaway, S. R. (1940). A Multiphasic Personality Schedule (Minnesota): II. A differential study of hypochondriasis. *Journal of Psychology, 10*, 255-268.

McKinley, J. C., & Hathaway, S. R. (1944). The MMPI: V. Hysteria, hypomania and psychopathic deviate. *Journal of Applied Psychology, 28*, 153-174.

McKinley, J. C., Hathaway, S. R., & Meehl, P. E. (1948). The MMPI: VI. The K scale. *Journal of Consulting Psychology, 12*, 20-31.

McLean, J. E., Kaufman, A. S., & Reynolds, C. R. (1989). Base rates of WAIS-R subtest scatter as a guide for clinical and neuropsychological assessment. *Journal of Clinical Psychology, 45*, 919-925.

McMordie, W. R. (1988). Twenty-year follow-up of the prevailing opinion on the posttraumatic or postconcussional syndrome. *The Clinical Neuropsychologyst, 2*, 198-212.

McMullin, R. E. (1986). *Handbook of cognitive therapy techniques*. New York: Norton.

McNemar, Q. (1964). Lost: Our intelligence? Why? *American Psychologist, 19*, 871-882.

McNulty, J., Graham, J., Ben-Porath, Y., & Stein, L. (1997). Comparative validity of MMPI-2 scores of African American and caucasian mental health center clients. *Psychological Assessment, 9*, 464-470.

Mecham, R. C., McCormick, E., & Jeanneret, P. (1977). *Technical manual for the Position Analysis Questionnaire* (PAQ) (System II). West Lafayette, IN: Purdue University Book Store.

Medina, N., & Neill, D. M. (1990). *Fallout from the testing explosion*. Cambridge, MA: Fairtest.

Meehl, P. E. (1954). *Clinical versus statistical prediction*. Minneapolis, MN: University of Minnesota Press.

Meehl, P. E. (1965). Seer over sign: The first good example. *Journal of Experimental Research in Personality, 1*, 29-32.

Meehl, P. E. (1986). Causes and effects of my disturbing little book. *Journal of Personality Assessment, 50*, 370-375.

Megargee, E. (1972). *The California Psychological Inventory handbook*. San Francisco: Jossey-Bass.

Meichenbaum, D. (1977). *Cognitive-behavior modification: An integrative approach*. New York: Plenum.

Meier, S. T. (1984). The construct validity of burnout. *Journal of Occupational Psychology, 57*, 211-219.

Meier, V. J., & Hope, D. A. (1998). Assessment of social skills. In A. S. Bellack & M. Hersen (eds.), *Behavioral assessment: A practical handbook* (4th ed.). Boston: Allyn & Bacon.

Melchert, T. P. (1998). Support for the validity of the Graduate Record Examination. *American Psychologyst, 53*, 573-574.

Melton, G. B. (1985). Review of Law School Admissions Test. In J. V. Mitchell, Jr. (ed.), *The ninth mental measurements yearbook*. Lincoln, NE: The Buros Institute of Mental Measurements.

Melton, G. B. (1995). Review of the Ackerman-Schoendorf Scales for Parent Evaluation of Custody. *The Twelfth Mental Measurements Yearbook*. Lincoln, NE: University of Nebraska Press.

Melton, G. B., Petrila, J., Poythress, N., & Slobogin, C. (1998). *Psychological evaluation for the courts* (2nd ed.). New York: Guilford.

Menzies, R. J., Webster, C., McMain, S., Staley, S., and Scaglione, R. (1994). The dimensions of dangerousness revisited: Assessing forensic predictions about violence. *Law and Human Behavior, 18*, 1-28.

Mercer, J. R., & Lewis, J. F. (1978). *System of Multicultural Pluralistic Assessment*. San Antonio, TX: The Psychological Corporation.

Merenda, P. F. (1985). Comrey Personality Scales. In D. J. Keyser & R. C. Sweetland (eds.), *Test critiques* (vol. 4). Kansas City, MO: Test Corporation of America.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*, 1012-1027.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

Messick, S., & Jungeblut, A. (1981). Time and method in coaching for the SAT. *Psychological Bulletin, 89*, 191-216.

Meyer, G. J. (1997). Assessing reliability: Critical corrections for a critical examination of the Rosrschach Comprehensive System. *Psychological Assessment, 9*, 480-489.

Mickley, J. (1990). *Spiritual well-being, religiousness, and hope: Some relationships in a sample of women with breast cancer*. Unpublished master's thesis, University of Maryland, School of Nursing, College Park, MD.

Miele, F. (1979). Cultural bias in the WISC. *Intelligence, 3*, 149-164.

Miller, L. K. (1989). *Musical savants: Exceptional skill in the mentally retarded.* Hillsdale, NJ: Erlbaum.

Miller, L. T., & Lee, C. (1993). Construct validation of the Peabody Picture Vocabulary Test-Revised: A structural equation model of the acquisition order of words. *Psychological Assessment, 5*, 438-441.

Miller, T. R. (1991). Personality: A clinician's experience. *Journal of Personality Assessment, 57*, 415-433.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (ed.), *Educational measurement* (3rd ed.). New York: ACE/Macmillan.

Millon, T. (1969). *Modern psychopathology: A biosocial approach to maladaptive learning and functioning*. Philadelphia: Saunders.

Millon, T. (1981). *Disorders of personality: DSM-III, Axis II*. New York: Wiley.

Millon, T. (1983). *Millon Clinical Multiaxial Inventory manual* (2nd ed.). Minneapolis, MN: National Computer Systems.

Millon, T. (1986). A theoretical derivation of pathological personalities. In T. Millon & G. Klerman (eds.), *Contemporary directions in psychopathology: Toward the DSM-IV*. New York: Guilford.

Millon, T. (1987). *Manual for the Millon Clinical Multiaxial Inventory-II (MCMI-II)* (2nd ed.). Minneapolis, MN: National Computer Systems.

Millon, T. (1994). *Manual for the Millon Clinical Multiaxial Inventory-III (MCMI-III)* (3rd ed.). Minneapolis, MN: National Computer Systems.

Millon, T., & Davis, R. (1996). The Millon Clinical Multiaxial Inventory-III (MCMI-III). In C. S. Newmark (ed.), *Major psychological assessment instruments* (2nd ed.). Boston: Allyn & Bacon.

Mills, C. N., & Stocking, M. L. (1996). Practical issues in large-scale computerized adaptive testing. *Applied Psychological Measurement, 9*, 287-304.

Mills, C., & Tissot, S. (1995). Identifying academic potential in students from underrepresented populations: Is using the Ravens Progressive Matrices a good idea? *Gifted Child Quarterly, 39*, 209-217.

Milmoe, S., Novey, M. S., Kagan, J., & Rosenthal, R. (1974). The mother's voice: postdictor of aspects of her baby's behavior. In S. Weitz (ed.), *Nonverbal communication: Readings with commentary*. New York: Oxford University Press.

Milmoe, S., Rosenthal, R., Blane, H. T., Chafetz, M. E., & Wolf I. (1967). The doctor's voice: Postdictor of successful referral of alcoholic patients. *Journal of Abnormal of Psychology, 72*, 78-84.

Milner, B. (1958). Psychological defects produced by temporal lobe excision. *Research Publications of the Association for Research in Nervous and Mental Disease, 38*, 244-257.

Milner, B., Corkin, S., & Teuber, H-L. (1968). Further analysis of the hippocampal amnesic syndrome: 14-year follow-up of H. M. *Neuropsychologia, 6*, 215-234.

Miner, J. B. (1964). *Scoring guide for the Miner Sentence Completion Scale*. New York: Springer.

Miner, M. G., & Miner, J. B. (1978). *Employee selection within the law*. Washington, DC: Bureau of National Affairs.

Mirsen, T., & Hachinski, V. (1988). Epidemiology and classification of vascular and multi-infarct dementia. In J. S. Meyer, J. Marshall, H. Lechner, & J. Toole (eds.), *Vascular and multi-infarct dementia*. Mount Kisco, NY: Futura Publishing.

Mischel, W. (1968). *Personality and assessment*. New York: Wiley.

Mishkin, M., & Appenzeller, T. (1987). The anatomy of memory. *Scientific American, 256*, 80-89.

Mishra, S. P., & Brown, K. (1983). The comparability of WAIS and WAIS-R IQs and subtest scores. *Journal of Clinical Psychology, 39*, 754-757.

Mitchell, J. L., & McCormick, E. J. (1979). *Development of the PMPQ: A structural job analysis questionnaire for the study of professional and managerial positions*. (Report No. 1). West Lafayette, IN: Occupational Research Center, Department of Psychological Studies, Purdue University.

Mitchell, J. V. (ed.). (1985). *The Ninth Mental Measurements Yearbook*. Lincoln, NE: University of Nebraska Press.

Mitchell, T. W., & Klimoski, R. J. (1986). Estimating the validity of cross-validity estimation. *Journal of Applied Psychology, 71*, 311-317.

Mittenberg, W., Azrin, R., Millsaps, C., & Heilbronner, R. (1993). Identification of malingered head injury on the Wechsler Memory Scale-Revised. *Psychological Assesment, 5*, 34-40.

Moberg, D. O. (1971). *Spiritual well-being: Background and issues*. Washington, DC: White House Conference on Aging.

Moore, E. G. J. (1986). Family socialization and the IQ-test performance of traditionally and transracially adopted children. *Developmental Psychology, 22*, 317-326.

Moore, W. P. (1994). The devaluation of standardized testing: One district's response. *Applied Measurement in Education, 7*, 343-368.

Moreland, K. L. (1985). Review of the Defining Issues Test. *Ninth mental measurements yearbook*. Lincoln, NE: University of Nebraska Press.

Moreland, K. L. (1992). Computer-assisted psychological assessment. In M. Zeidner and R. Most (eds.), *Psychological testing: An inside view*. Palo Alto, CA: Consulting Psychologists Press.

Morgan, C. D., & Murray, H. A. (1935). A method for investigating phantasies: The Thematic Apperception Test. *Archives of Neurology and Psychiatry, 34*, 289-306.

Morgan, S., & Wheelock, J. (1992). Digit Symbol and Symbol Digit Modalities Tests: Are they directly interchangeable? *Neuropsychology, 6*, 327-330.

Morgeson, F, P., & Campion, M. A. (1997). Social and cognitive science sources of potential inaccuracy in job analysis. *Journal of Applied Psychology, 82*, 627-655.

Morris, L. W., Davis, M. A., & Hutchings, C. A. (1981). Cognitive and emotional components of anxiety: Literature review and a revised worry-emotionality scale. *Journal of Educational Psychology, 73*, 541-555.

Morrison, M. W., Gregory, R. J., & Paul, J. J. (1979). Reliability of the Finger Tapping Test and a note on sex differences. *Perceptual and Motor Skills, 48*, 139-142.

Morrison, T., & Morrison, M. (1995). A meta-analytic assessment of the predictive validity of the quantitative and verbal components of the Graduate Record Examination with graduate grade point average representing the criterion of graduate success. *Educational and Psychological Measurement, 55*, 309-316.

Moruzzi, G., & Magoun, H. W. (1949). Brain stem and reticular formation and activation of the EEG. *Electroencephalography and Clinical N europhysiology, 1*, 455-473.

Moses, J. A., & Golden, C. J. (1979). Cross validation of the discriminative effectiveness of the standardized Luria Neuropsychological Battery. *International Journal of Neuroscience, 9*, 149-155.

Moskowitz, D. S. (1982). Coherence and cross-situational generality in personality: A new analysis of old problems. *Journal of Personality and Social Psychology, 43*, 754-768.

Motowidlo, S. J., Carter, G., Dunnette, M., Tippins, N., Werner, S., Burnett, J., & Vaughan, M. (1992). Studies of the structured behavioral interview. *Journal of Applied Psychology, 77*, 571-587.

Motta, R. W., Little, S., & Tobin, M. (1993). The use and abuse of human figure drawings. *School Psychology Quarterly, 8*, 162-169.

Mountain, M., & Snow, W. (1993). Wisconsin Card Sorting Test as a measure of frontal pathology: A review. *Clinical Neuropsychologist, 7*, 108-118.

Muchinsky, P. (1985). Review of Values Inventory. *Ninth Mental Measurements Yearbook*. Lincoln, NE: University of Nebraska Press.

Muchinsky, P. M. (1990). *Psychology applied to work: An introduction to industrial and organizational psychology* (3rd ed.). Chicago: Dorsey Press.

Mumford, M. D., & Owens, W. A. (1987). Methodology review: principles, procedures, and findings in the application of background data measures. *Applied Psychological Measurement, 11*, 1-31.

Mundfrom, D., Bradley, R., & Whiteside, L. (1993). A factor analytic study of the Infant-Toddler and Early Childhood versions of the HOME Inventory. *Educational and Psychological Measurement, 53*, 479-489.

Mungas, D. M., Trontel, E. H., & Weingardner, J. (1981). Multivariable-multimethod analysis of the dimensions of interpersonal behavior. *Journal of Research in Personality, 15*, 107-121.

Murphy, K. R. (1984). Review of Armed Services Vocational Aptitude Battery. In D. Keyser & R. Sweetland (eds.), *Test critiques* (vol. I). Kansas City, MO: Test Corporation of America.

Murphy, K. R. (1992). Review of TONI-2. *The Eleventh Mental Measurements Yearbook*. Lincoln, NE: University of Nebraska Press.

Murphy, K. R. (ed.). (1996). *Individual differences and behavior in organizations*. San Francisco: Jossey-Bass.

Murphy, K. R., & Davidshofer, C. O. (1988). *Psychological testing*. Englewood Cliffs, NJ: Prentice-Hall.

Murphy, K. R., & Pardaffy, V. A. (1989). Bias in behaviorally anchored rating scales: Global or scale-specific? *Journal of Applied Psychology, 74*, 343-346.

Murphy, K. R., Jako, R., & Anhalt, R. (1993). Nature and consequences of halo error: A critical analysis. *Journal of Applied Psychology, 78*, 218-225.

Murray, H. A. (1938). *Explorations in personality*. New York: Oxford University Press.

Murray, H. A. (1943). *Thematic Apperception Test—Manual*. Cambridge: Harvard University Press.

Museum of Modern Art (1955). *The family of man*. New York: Maco Magazine Corporation.

Myers, I. B., & McCaulley, M. H. (1985). *Manual: A guide to the development and use of the Myers-Briggs Type Indicator*. Palo Alto, CA: Consulting Psychologists Press.

Naglieri, J. A. (1981). Concurrent validity of the Revised Peabody Picture Vocabulary Test. *Psychology in the Schools, 18*, 286-289.

Naglieri, J. A. (1985). Use of the WISC-R and K-ABC with learning disabled, borderline mentally retarded, and normal children. *Psychology in the Schools, 22*, 133-141.

Naglieri, J. A. (1988). *Draw A Person: A quantitative scoring system*. San Antonio, TX: The Psychological Corporation.

Naglieri, J. A., & Pfeiffer, S. (1983). Stability, concurrent and predictive validity of the PPVT-R. *Journal of Clinical Psychology, 39*, 965-967.

Naglieri, J. A., & Pfeiffer, S. (1992). Performance of disruptive behavior disordered and normal samples

on the Draw A Person: Screening Procedure for Emotional Disturbance. *Psychological Assessment, 4*, 156-159.

Naglieri, J. A., & Yazzie, C. (1983). Comparison of the WISC-R and PPVT-R with Navajo children. *Journal of Clinical Psychology, 39*, 598-600.

Naglieri, J. A., McNeish, T., & Bardos, A. (1991). *Draw-A-Person: Screening Procedure for Emotional Disturbance*. Austin, TX: ProEd.

Nass, R. (1992). Developmental dyslexia: An update. *Pediatrics in Review, 13*, 231-235.

National Association of School Psychologists. (1992). *Principles for professional ethics*. Silver Spring, MD: Author.

National Joint Committee on Learning Disabilities. (1988). A position paper of the National Trust Committee on Learning Disabilities. *Journal of Learning Disabilities, 21*, 53-55.

Naugle, R. I., Chelune, G., & Tucker, G. (1993). Validity of the Kaufman Brief Intelligence Test. *Psychological Assessment, 5*, 182-186.

Nauta, W. J. H. (1971). The problem of the frontal lobe. *Journal of Psychiatric Research, 8*, 167-187.

Naveh-Benjamin, M., McKeachie, W. J., & Lin, Y. (1987). Two types of test-anxious students: Support for an information processing model. *Journal of Educational Psychology, 79*, 131-136.

Needleman, H. L., Gunnoe, C., Leviton, A., Reed, R., Peresie, H., Maher, C., & Barrett, P. (1979). Deficits in psychologic and classroom performance of children with elevated dentine lead levels. *The New England Journal of Medicine, 300*, 689-695.

Needleman, H. L., Schell, A., Bellinger, D., Leviton, A., & Allred, E. (1990). The long-term effects of exposure to low doses of lead in childhood. *New England Journal of Medicine, 322*, 83-88.

Neisser, U. (ed.). (1998). *The rising curve: Long-term gains in IQ and related measures*. Washington, DC: American Psychological Association.

Neisser, U., Boodoo, G., & Bouchard, T., and others (1996). Intelligence: Knowns and unknowns. *American Psychlogist, 51*, 77-101.

Nellis, L., & Gridley, B. (1994). Review of the Bayley Scales of Infant Development-Second Edition. *Journal of School Psychology, 32*, 201-209.

Nester, M. A. (1994). Psychometric testing and reasonable accommodation for persons with disabilities. In S. M. Bruyere & J. O'Keeffe (eds.), *Implications of the Americans with Disabilities Act for psychology*. New York: Springer.

Netter, B., & Viglione, D., Jr. (1994). An empirical study of malingering schizophrenia on the Rorschach. *Journal of Personality Assessment, 62*, 45-57.

Nevo, B. (1985). Face validity revisited. *Journal of Educational Measurement, 22*, 287-293.

Nevo, B. (1992). Examinee feedback: Practical guidelines. In M. Zeidner & R. Most (eds.), *Psychological testing: An inside view*. Palo Alto, CA: Consulting Psychologists Press.

Newcomer, P. (1990). *Diagnostic Achievement Battery manual* (2nd ed.) Austin, TX: ProEd.

Newland, T. E. (1971). *Blind Learning Aptitude Test*. Champaign, IL: University of Illinois Press.

Nicholson, R. A., Robertson, H., Johnson, W., & Jensen, G. (1988). A comparison of instruments for assessing competency to stand trial. *Law and Human Behavior, 12*, 313-321.

Nielsen, G. (1962). *Studies in self confrontation*. Copenhage: Monksgaard.

Nihira, K. (1985). Assessment of mentally retarded individuals. In B. B. Wolman (ed.), *Handbook of intelligence: Theories, measurements, and applications*. New York: Wiley.

Nihira, K., Leland, H., & Lambert, N. (1993). *Adaptive Behavior Scale-Residential and Community: 2nd Ed*. Washington, DC: American Association on Mental Retardation.

Nijenhuis, J., & van der Flier, H. (1997). Comparability of GATB scores for immigrants and majority group members: Some Dutch findings. *Journal of Applied Psychology, 82*, 675-687.

Nisan, M., & Kohlberg, L. (1982). Universality and cross-cultural variation in moral development: A longitudinal and cross-sectional study in Turkey. *Child Development, 53*, 865-876.

Nitko, A. J. (1989). Review of Metropolitan Achievement Test, Sixth Edition. *The Tenth Mental Measurements Yearbook*. Lincoln, NE: University of Nebraska Press.

Nolan, R. F., Watlington, D. K., & Willson, V. L. (1989). Gifted and nongifted race and gender effects on item functioning on the Kaufman Assessment Battery for Children. *Journal of Clinical Psychology, 45*, 645-650.

Nottingham, E. J., & Mattson, R. E. (1981). A validation study of the Competency Screening Test. *Law and Human Behavior, 5*, 329-335.

Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika, 32*, 1-13.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Nuttall, E. V., Romero, I., & Kalesnik, J. (1992). *Assessing and screening preschoolers: Psychological and educational dimensions*. Boston, MA: Allyn & Bacon.

O'Neill, J., Jacobson, S., & Jacobson, J. (1994). Evidence of observer reliability for the Fagan Test of Infant Intelligence (FTII). *Infant Behavior and Development, 17*, 465-469.

Oakland, T. D., & Dowling, L. (1983). The Draw-A-Person test: Validity properties for nonbiased assessment. *Learning Disability Quarterly, 6,* 526-534.

O'Bannon, R. M., Goldinger, L. A., & Appleby, G. S. (1989). *Honesty and integrity testing.* Atlanta, GA: Applied Information Resources.

Obrzut, A., Obrzut, J. E., & Shaw, D. (1984). Construct validity of the Kaufman Assessment Battery for Children with learning disabled and mentally retarded. *Psychology in the Schools, 21,* 417-424.

Oehler-Stinnett, J. (1992). Review of Conners' Rating Scales. *Eleventh Mental Measurements Yearbook.* Lincoln, NE: University of Nebraska Press.

Oles, H. J., & Davis, G. D. (1977). Publishers violate APA standards on test distribution. *Psychological Reports, 41,* 713-714.

Ollendick, T. H. (1983). Reliability and validity of the Revised Fear Survey Schedule for Children (FSSC-R). *Behavior Research and Therapy, 21,* 685-692.

Olson, H. C. (1994). Fetal alcohol syndrome. In R. J. Sternberg (ed.), *Encyclopedia of human intelligence.* New York: Macmillan.

Ones, D. S., & Viswesvaran, C. (1998). Gender, age, and race differences on overt integrity tests: Results across four large-scale job applicant data sets. *Journal of Applied Psychology, 83,* 35-42.

Ones, D. S., Viswesvaran, C., & Schmidt, F. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology, 78,* 679-703.

Ornberg, B., & Zalewski, C. (1994). Assessment of adolescents with the Rorschach: A critical review. *Assessment, 1,* 209-217.

Orpen, C. (1983). The development and validation of an adjective check-list measure of managerial need for achievement. *Psychology: A Quarterly Journal of Human Behavior, 20,* 38-42.

Osipow, S. H. (1980). *Manual for the Career Decision Scale.* Columbus, OH: Marathon Consulting and Press.

OSS Assessment Staff. (1948). *Assessment of men: Selection of personnel for the Office of Strategic Services.* New York: Rinehart.

Otis, A. S. (1918). An absolute point scale for the group measure of intelligence. *Journal of Educational Psychology, 9,* 238-261, 333-348.

Overholser, J. C. (1990). Retest reliability of the Millon Clinical Multiaxial Inventory. *Journal of Personality Assessment, 55,* 202-208.

Owens, W. A. (1976). Background data. In M. D. Dunnette (ed.), *Handbook of industrial and organizational psychology.* Chicago: Rand McNally.

Ownby, R. L. (1991). *Psychological reports: A guide to report writing in professional psychology* (2nd ed.). Brandon, VT: Clinical Psychology Publishing Co.

Padilla, A., & Medina, A. (1996). Cross-cultural sensitivity in assessment: Using tests in culturally appropriate ways. In L. A. Suzuki, P. J. Meller, & J. G. Ponterotto (eds.), *Handbook of multicultural assessment: Clinical, psychological, and educational applications.* Englewood Cliffs, NJ: Prentice-Hall.

Paget, K. D. (1990). Assessment of intellectual competence in preschool-age children: Conceptual issues and challenges. In C. R. Reynolds and R. W. Kamphaus (eds.). *Handbook of psychological and educational assessment of children: Intelligence and achievement.* New York: Guilford.

Paloutzian, R. F., & Ellison, C. W. (1982). Loneliness, spiritual well-being and the quality of life. In L. A. Peplau & D. Perlman (eds.), *Loneliness: A sourcebook of current theory, research and therapy.* New York: Wiley.

Panigua, F. (1994). *Assessing and treating culturally diverse clients: A practical guide.* Thousand Oaks, CA: Sage.

Papez, J. W. (1937). A proposed mechanism of emotion. *Archives of Neurology and Psychiatry, 38,* 724-744.

Parker, K. C., & Atkinson, L. (1994). Factor space of the Wechsler Intelligence Scale for Children-Third Edition: Critical thoughts and recommendations. *Psychological Assessment, 6,* 201-208.

Pascal, G. R., & Suttell, B. J. (1951). *The Bender-Gestalt Test.* New York: Grune & Stratton.

Patrick, J. (1988). Concordance of the MCMI and the MMPI in the diagnosis of three DSM-III Axis I disorders. *Journal of Clinical Psychology, 44,* 186-190.

Patterson, C. (1980). An alternative perspective—lead pollution in the human environment. In *Lead in the human environment.* Washington, DC: National Academy of Sciences.

Patton, J. R., Payne, J. S., & Beirne-Smith, M. (1986). *Mental retardation* (2nd ed.). Columbus, OH: Merrill.

Pauker, J. D. (1976). A quick-scoring system for the Bender-Gestalt: Interrater reliability and scoring validity. *Journal of Clinical Psychology, 32,* 86-89.

Paul, G. L. (ed.). (1986). *Assessment in residential treatment settings.* Champaign, IL: Research Press.

Paulhus, D., Fridhandler, B., & Hayes, S. (1997). Psychological defense: Contemporary theory and research. In R. Hogan, J. Johnson, & S. Briggs (eds.), *Handbook of personality psychology.* San Diego: Academic Press.

Paulman, R. G., & Kennelly, K. J. (1984). Test anxiety and ineffective test taking: Different names, same construct? *Journal of Educational Psychology, 76,* 279-288.

Payne, A. F. (1928). *Sentence completions.* New York: New York Guidance Clinic.

Pearson, K. (1914, 1924, 1930ab). *The life, letters, and labours of Francis Galton* (Volumes I, II, III, IIIb). Cambridge: Cambridge University Press.

Pedersen, N. L., Plomin, R., Nesselroade, J., & McClearn, G. (1992). A quantitative genetic analysis of cognitive abilities during the second half of the life span. *Psychological Science, 3*, 346-353.

Penfield, W. (1958). Functional localization in temporal and deep sylvian areas. *Research Publication, Association of Nervous and Mental Disease, 36*, 210-217.

Penfield, W., & Evans, J. (1935). The frontal lobe in man: A clinical study of maximum removals. *Brain, 58*, 115-133.

Penfield, W., & Jasper, H. (1959). *Epilepsy and the functional anatomy of the human brain.* Boston: Little, Brown.

Perleth, C., Hofmann, U., Schauer, S., & Wernberger, H. (1994). Intelligence testing in a Bavarian school. *School Psychology International, 15*, 261-275.

Pervin, L. A. (1993). *Personality: Theory and research* (6th ed.). New York: Wiley.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling norming and equiating. In R. L. Linn (ed.), *Educational Measurement* (3rd ed.). New York: American Council on Education/ Macmillan.

Peterson, N. G., Mumford, M., Borman, W., and others (1995). *Development of prototype occupational information network (O*NET) content model* (vols. 1 & 2). Salt Lake City, UT: Utah Department of Employment Security.

Peterson, R. A. (1978). Review of the Rorschach. *The Eighth Mental Measurements Yearbook.* Lincoln, NE: University of Nebraska Press.

Phares, E. J. (1985). Incomplete sentences. In R. J. Corsini (ed.), *Encyclopedia of psychology* (vol. 2). New York: Wiley.

Phelps, L., & Ensor, A. (1986). Concurrent validity of the WISC-R using deaf norms and the Hiskey-Nebraska. *Psychology in the Schools, 23*, 138-141.

Phillips, S. E. (1994). High-stakes testing accommodations: Validity versus disabled rights. *Applied Measurement in Education, 7*, 93-120.

Piaget, J. (1926). *Judgment and reasoning in the child.* New York: Harcourt, Brace & World.

Piaget, J. (1932). *The moral judgment of the child.* London: Kegan Paul.

Piaget, J. (1952). *The origins of intelligence in children.* New York: International Universities Press.

Piaget, J. (1972). *The psychology of intelligence.* Totowa, NJ: Littlefield Adams.

Piedmont, R. L., & Weinstein, H. P. (1993). A psychometric evaluation of the new NEO-PIR Facet Scales for Agreeableness and Conscientiousness. *Journal of Personality Assessment, 60*, 302-318.

Piedmont, R. L., McCrae, R. R., & Costa, P. T., Jr. (1992). An assessment of the Edwards Personal Preference Schedule from the perspective of the five-factor model. *Journal of Personality Assessment, 58*, 67-78.

Piersma, H., & Boes, J. (1997). MCMI-III as a treatment outcome measure for psychiatric inpatients. *Journal of Clinical Psychology, 53*, 825-832.

Pintner, R. (1917). The mentality of the dependent child. *Journal of Educational Psychology, 8*, 220-238.

Pintner, R. (1921). Intelligence. In E. L. Thorndike (ed.), Intelligence and its measurement: A symposium. *Journal of Educational Psychology, 12*, 123-147, 195-216.

Pintner, R., & Paterson, D. G. (1917). *A scale of performance tests.* New York: Appleton.

Piotrowski, C. (1996). The status of Exner's Comprehensive System in contemporary research. *Perceptual and Motor Skills, 82*, 1341-1342.

Piotrowski, Z. A. (1964). A digital computer administration of inkblot test data. *Psychiatric Quarterly, 38*, 1-26.

Pirozzolo, F. J., Hansch, E., Mortimer, J., Webster, D., & Kuskowski, A. (1982). Dementia in Parkinson disease: A neuropsychological analysis. *Brain and Cognition, 1*, 71-83.

Plaisted, J. R., & Golden, C. J. (1982). Test-retest reliability of the clinical, factor and localization scales of the Luria-Nebraska Neuropsychological Battery. *International Journal of Neuroscience, 17*, 163-167.

Plaud, J. J., & Eifert, G. (eds.), (1998). *From behavior theory to behavior therapy.* Boston: Allyn & Bacon.

Pollack, R. H. (1971). Binet on perceptual-cognitive development or Piaget-come-lately. *Journal of the History of the Behavioral Sciences, 7*, 370-374.

Pollens, R., McBratnie, B., & Burton, P. (1988). Beyond cognition: Executive functions. *Cognitive Rehabilitation, 6*, 26-33.

Pope, K. S. (1992). Responsibilities in providing psychological test feedback to clients. *Psychological Assessment, 4*, 268-271.

Popham, W. J. (1978). *Criterion-referenced measurement.* Englewood Cliffs, NJ: Prentice-Hall.

Porch, B. E. (1967). *Porch Index of Communicative Ability.* Palo Alto, CA: Consulting Psychologists Press.

Porteus, S. D. (1915). Mental tests for the feebleminded: A new series. *Journal of Psycho-Asthenics, 19*, 200-213.

Porteus, S. D. (1931). *The psychology of a primitive people: A study of the Australian aborigine.* London: Edward Arnold & Co.

Porteus, S. D. (1965). *Porteus Maze Test. Fifty years' application.* Palo Alto, CA: Pacific Books.

Powers, D. E., & Swinton, S. S. (1984). Effects of self-study for coachable test item types. *Journal of Educational Psychology, 76*, 266-278.

Prewett, P. N. (1995). A comparison of two screening tests (the Matrix Analogies Test-Short Form and the Kaufman Brief Intelligence Test) with the WISC-III. *Psychological Assessment, 7*, 69-72.

Prewett, P. N., & Matavich, M. (1994). A comparison of referred students' performance on the WISC-III and the Stanford-Binet Intelligence Scale-Fourth Edition. *Journal of Psychoeducational Assessment, 12*, 42-48.

Prout, H., & Schwartz, J. (1984). Validity of the PPVT-R with mentally retarded adults. *Journal of Clinical Psychology, 40*, 584-587.

Psychological Corporation (1994). *WISC-III Writer manual*. San Antonio, TX: Author.

Pyle, W. H. (1913). *The examination of school children*. New York: Macmillan.

Quay, H., & Peterson, D. (1983). *Interim manual for the Revised Behavior Problem Checklist*. Unpublished manuscript, University of Miami.

Raichle, M. E. (1994). Visualizing the mind. *Scientific American*, April, 36-42.

Raju, N. S. (1992). Review of Iowa Tests of Basic Skills. *The Eleventh Mental Measurements Yearbook*. Lincoln, NE: University of Nebraska Press.

Ramey, C. T., & Ramey, S. (1998). Early intervention and early experience. *American Psychologist, 53*, 109-10.

Rasch, G. (1966). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: The Danish Institute for Educational Research.

Raskin, D. C. (1989). *Psychological methods in criminal investigation and evidence*. New York: Springer.

Rausch, H. R. (1985). Differences in cognitive function with left and right temporal lobe dysfunction. In D. F. Benson & E. Zaidel (eds.), *The dual brain: Hemispheric specialization in humans*. New York: Guilford.

Raven, J. C. (1938). *Progressive Matrices*. London: Lewis.

Raven, J. C. (1965). *The Coloured Pregressive Matrices Test*. London: Lewis.

Raven, J. C., & Summers, B. (1986). *Manual for Raven's Progressive Matrices and Vocabulary Scales—research supplement no. 3*. London: Lewis.

Raven, J. C., Court, J. H., & Raven, J. (1983). *Manual for Raven's Progressive Matrices and Vocabulary Scales (Section 3)—Standard Progressive Matrices* (1983 edition). London: Lewis.

Raven, J. C., Court, J. H., & Raven, J. (1986). *Manual for Raven's Progressive Matrices and Vocabulary Scales (Section 2)—Coloured Progressive Matrices* (1986 edition, with U.S. norms). London: Lewis.

Raven, J. C., Court, J. H., & Raven, J. (1992). *Standard Progressive Matrices. 1992 Edition*. Oxford: Oxford Psychologists Press.

Reddon, J. R., & Jackson, D. N. (1989). Readability of three adult personality tests: Basic Personality Inventory, Jackson Personality Inventory, and Personality Research Form-E. *Journal of Personality Assessment, 53*, 180-183.

Ree, M. J., & Carretta, T. R. (1994). Factor analysis of the ASVAB: Confirming a Vernon-like structure. *Educational and Psychological Measurement, 54*, 459-463.

Ree, M. J., & Carretta, T. R. (1995). Group differences in aptitude factor structure on the ASVAB. *Educational and Psychological Measurement, 55*, 268-277.

Ree, M. J., Earles, J., & Teachout, M. (1994). Predicting job performance: not much more than *g*. *Journal of Applied Psychology, 79*, 518-524.

Reed, M. T., & McCallum, R. S. (1995). Construct validity of the Universal Nonverbal Intelligence Test (UNIT). *Psychology in the Schools, 32*, 277-290.

Reeve, R. R., French, J. L., & Hunter, M. (1983). A validation of the Leiter International Performance Scale with kindergarten children. *Journal of Consulting and Clinical Psychology, 51*, 331-337.

Reeves, D., & Wedding, D. (1994). *The clinical assessment of memory: A practical guide*. New York: Springer.

Rehm, L. P. (1984). Self-management therapy for depression. *Advances in Behavior Research and Therapy, 6*, 83-98.

Rehm, L. P., Kornblith, S. J., O'Hara, M. W., and others (1981). An evaluation of major components in a self-control therapy program for depression. *Behavior Modification, 5*, 459-490.

Reid, D. B., & Kelly, M. P. (1993). Wechsler Memory Scale-Revised in closed head injury. *Journal of Clinical Psychology, 49*, 245-254.

Reilly, R. R., & Chao, G. T. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology, 35*, 1-63.

Reitan, R. M. (1984). *Aphasia and sensory-perceptual deficits in adults*. Tucson, AZ: Neuropsychology Press.

Reitan, R. M. (1985). *Aphasia and sensory-perceptual deficits in children*. Tucson, AZ: Neuropsychology Press.

Reitan, R. M., & Wolfson, D. (1985). *The Halstead-Reitan Neuropsychological Test Battery: Theory and clinical interpretation*. Tucson, AZ: Neuropsychology Press.

Reitan, R. M., & Wolfson, D. (1993). *The Halstead-Reitan Neuropsychological Test Battery: Theory and clinical interpretation* (2nd ed.). Tucson, AZ: Neuropsychology Press.

Renzulli, J. S. (1978). What makes giftedness? Reexamining a definition. *Phi Delta Kappan, 60*, 180-184.

Renzulli, J. S. (1986). The three-ring conception of giftedness: A developmental model for creative productivity. In R. J. Sternberg & J. E. Davidson (eds.), *Conceptions of giftedness*. Cambridge: Cambridge University Press.

Reschly, D. J. (1978). WISC-R factor structures among Anglos, Blacks, Chicanos, and Native-American Papagos. *Journal of Consulting and Clinical Psychology, 3*, 417-422.

Reschly, D. J. (1990). Adaptive behavior. In A. Thomas & J. Grimes (eds.), *Best practices in school psychology* (2nd ed.). Washington, DC: National Association of School Psychologists.

Rest, J. R. (1979). *The Defining Issues Test: Manual*. Minneapolis: University of Minnesota Press.

Rest, J. R. (1986). Moral research methodology. In S. Modgil & C. Modgil (eds.), *Lawrence Kohlberg: Consensus and controversy*. Philadelphia, PA: Taylor & Francis.

Rest, J. R., & Thoma, S. J. (1985). Relation of moral judgment to formal education. *Developmental Psychology, 21*, 709-714.

Rest, J. R., Thoma, S., Narvaez, D., & Bebeau, M. (1997). Alchemy and beyond: Indexing the Defining Issues Test. *Journal of Educational Psychology, 89*, 498-507.

Rey, A. (1964). *L'examen clinique en psychologie*. Paris: Presses Universitaires de France.

Reynolds, C. R. (1982). The problem of bias in psychological assessment. In C. R. Reynolds & T. B. Gutkin (eds.), *The handbook of school psychology*. New York: Wiley.

Reynolds, C. R. (1992). Review of the Millon Clinical Multiaxial Inventory-II. *Eleventh Mental Measurements Yearbook*. Lincoln, NE: University of Nebraska Press.

Reynolds, C. R. (1994a). Bias in testing. In R. J. Sternberg (ed.), *Encyclopedia of human intelligence*. New York: Macmillan.

Reynolds, C. R. (1994b). Kaufman Assessment Battery for Children (K-ABC). In R. J. Sternberg (ed.), *Encyclopedia of human intelligence*. New York: Macmillan.

Reynolds, C. R., & Brown, R. T. (1984a). Bias in mental testing: An introduction to the issues. In Reynolds, C. R., & Brown, R. T. (eds.), *Perspectives on bias in mental testing*. New York: Plenum.

Reynolds, C. R., & Brown, R. T. (eds.). (1984b). *Perspectives on bias in mental testing*. New York: Plenum.

Reynolds, C. R., & Kamphaus, R. (1992). *Behavior assessment system for children*. Circle Pines, MN: American Guidance Service.

Reynolds, C. R., Chastain, R. L., Kaufman, A. S., & McLean, J. E. (1987). Demographic characteristics and IQ among adults: Analysis of the WAIS-R standardization sample as a function of the stratification variables. *Journal of School Psychology, 25*, 323-342.

Reynolds, C. R., Willson, V. L., and Chatman, S. (1984). Relationships between age and raw score increases on the Kaufman-Assessment Battery for Children. *Psychology in the Schools, 21*, 19-24.

Rhodes, L., Bayley, N., and Yow, B. (1983). *Manual supplement: Bayley Scales of Infant Development*. San Antonio, TX: Psychological Corporation.

Riccio, C. A., Cohen, M., Hall, J., & Ross, C. (1997). The third and fourth factors of the WISC-III: What they don't measure. *Journal of Psychoeducational Assessment, 15*, 27-39.

Ricciuti, H. N. (1994). Infant tests as measures of early competence. In R. J. Sternberg (ed.), *Encyclopedia of human intelligence*. New York: Macmillan.

Richards, P. S. (1991). The relation between conservative religious ideology and principled moral reasoning: A review. *Review of Religious Research, 32*, 359-368.

Richards, P. S., & Bergin, A. E. (1997). *A spiritual strategy for counseling and psychotherapy*. Washington, DC: American Psychological Association.

Richards, P. S., & Davison, M. L. (1992). Religious bias in moral development research: A psychometric investigation. *Journal for the Scientific Study of Religion, 31*, 467-485.

Rieber, R. W. (ed.). (1980). *Wilhelm Wundt and the making of a scientific psychology*. New York: Plenum Press.

Rinas, J., & Clyne-Jackson, S. (1988). *Professional conduct and legal concerns in mental health practice*. Norwalk, CT: Appleton & Lang.

Ritzler, B. A. (1993). TEMAS (Tell-Me-A-Story): Review. *Journal of Psychoeducational Assessment, 11*, 381-389.

Ritzler, B. A., Sharkey, K. J., & Chudy, J. (1980). A comprehensive projective alternative to the TAT. *Journal of Personality Assessment, 44*, 358-362.

Robertson, I. H., Ward, T., Ridgeway, V., & Nimmo-Smith, I. (1994). *Test of Everyday Attention (TEA)*. Gaylord, MI: National Rehabilitation Services.

Robertson, I. H., Ward, T., Ridgeway, V., & Nimmo-Smith, I. (1996). The structure of normal human attention: The Test of Everyday Attention. *Journal of the International Neuropsychological Society, 2*, 525-534.

Robinson, R. (1950). *Definition*. Oxford: Oxford University Press.

Roecker, C., House, A., & Graybill, D. (1992-93). Luria-Nebraska Neuropsychological Battery: High rates of false positives for geriatric subjects. *Current Psychology: Research and Practice, 11*, 354-359.

Rogers, B. (1989). Review of Metropolitan Achievement Test, Sixth Edition. *The Tenth Mental Measurements Yearbook*. Lincoln, NE: University of Nebraska Press.

Rogers, B. G. (1992). Review of GED. *The Eleventh Mental Measurements Yearbook*. Lincoln, NE: University of Nebraska Press.

Rogers, C. R. (1951). *Client-centered therapy: Its current practice, implications, and theory*. Boston: Houghton Mifflin.

Rogers, C. R. (1961). *On becoming a person: A therapyst's view of psychotherapy*. Boston: Houghton Mifflin.

Rogers, C. R. (1980). *A way of being*. Boston: Houghton Mifflin.

Rogers, C. R., & Dymond, R. F. (eds.). (1954). *Psychotherapy and personality change: Coordinated research studies in the client-centered approach*. Chicago: University of Chicago Press.

Rogers, M. R. (1998). Psychoeducational assessment of culturally and linguistically diverse children and youth. In H. B. Vance (ed.), *Psychological assessment of children: Best practices for schools and clinical settings* (2nd ed.). New York: Wiley.

Rogers, R. (1984). *Rogers Criminal Responsibility Assessment Scales*. Odessa, FL: Psychological Assessment Resources.

Rogers, R. (1986). *Conducting insanity evaluations*. Odessa, FL: Psychological Assessment Resources.

Rogers, R., Bagby, M., & Dickens, S. (1992). *Structured Interview of Reported Symptoms (SIRS) manual*. Odessa, FL: Psychological Assessment Resources.

Rogers, R., Gillis, J., Dickens, S., & Bagby, R. (1991). Standardized assessment of malingering: Validation of the SIRS. *Psychological Assessment, 3*, 89-96.

Rogers, R., Kropp, P., Bagby, R., & Dickens, S. (1992). Faking specific disorders: A study of the SIRS. *Journal of Clinical Psychology, 48*, 643-648.

Rogoff, B. (1984). What are the interrelations among the three subtheories of Sternberg's triarchic theory of intelligence? *Behavioral and Brain Sciences, 7*, 300-301.

Roid, G. H., & Johnson, W. B. (1998). Computer assisted psychological assessment. In A. S. Bellack & M. Hersen (eds.), *Comprehensive clinical psychology* (vol. 4). Amsterdam: Elsevier.

Roid, G., & Miller, L. (1997). *Leiter-R Manual*. Wood Dale, IL: Soelting Co.

Roid, G., & Worrall, W. (1997). Replication of the Wechsler Intelligence Scale for Children-Third Edition four-factor model in the Canadian normative sample. *Psychological Assessment, 9*, 512-515.

Rokeach, M. (1968). *Beliefs, attitudes, and values*. San Francisco: Jossey-Bass.

Rokeach, M. (1973). *The nature of human values*. New York: Free Press.

Rokeach, M. (1980). Some unresolved issues in theories of beliefs, attitudes, and values. In H. E. Howe, Jr., & M. M. Page (eds.), *Nebraska symposium on motivation*. Lincoln. NE: University of Nebraska Press.

Ronan, G. F., Colavito, V., & Hammontree, S. (1993). Personal Problem-Solving System for scoring TAT responses: Preliminary validity and reliability data. *Journal of Personality Assessment, 61*, 28-40.

Rorschach, H. (1921). *Psychodiagnostik*. Berne: Birchen.

Rosenberg, S., Ryan, J., & Prifitera, A. (1984). Rey Auditory-Verbal Learning Test performance of patients with and without memory impairment. *Journal of Clinical Psychology, 40*, 785-787.

Rosenthal, R (1967). Covert communication in the psychological experiment. *Psychological Bulletin, 67*, 356-367.

Rosenthal, R., Hall, J. A., DiMatteo, M. R., Rogers, P. L., & Archer, D. (1979). *Sensitivity to nonverbal communication: The PONS test*. Baltimore, MD: Johns Hopkins University Press.

Rosenzweig, S. (1944). An outline of frustration theory. In J. McV. Hunt (ed.), *Personality and the behavior disorders* (vol. 1). New York: Ronald Press.

Rosenzweig, S. (1977). *Manual for the Children's Form of the Rosenzweig Picture-Frustration (P-F) Study*. St. Louis: Rana House.

Rosenzweig, S. (1978a). *The Rosenzweig Picture-Frustration (P-F) Study: Basic manual*. St. Louis: Rana House.

Rosenzweig, S. (1978b). *Aggresive behavior and the Rosenzweig Picture-Frustration Study*. New York: Praeger.

Rosenzweig, S., & Adelman, S. (1977). Construct validity of the Picture-Frustration Study. *Journal of Personality Assessment, 41*, 578-588.

Ross, S. M., Gottfredson, D. K., Christensen, P., & Weaver, R. (1986). Cognitive self-statements in depression: Findings across clinical populations. *Cognitive Therapy and Research, 10*, 159-166.

Rosvold, H. E., Mirsky, A. E., Sarason, I., and others (1956). A continuous performance test of brain damage. *Journal of Consulting Psychology, 20*, 343-350.

Rothermel, R. D., & Lovell, M. R. (1985). Personality Inventory for Children. In D. J. Keyser & R. C. Sweetland (eds.), *Test critiques* (vol. 2). Kansas City, MO: Test Corporation of America.

Rothstein, H. R., Schmidt, F. L., Erwin, F. W., Owens, W. A., & Sparks, C. P. (1990). Biographical data in employment selection: Can validities be made generalizable? *Journal of Applied Psychology, 75*, 175-184.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs, 80* (Whole No. 609).

Rotter, J. B. (1972). Beliefs, social attitudes, and behavior: A social learning analysis. In J. B. Rotter, J. Chances, & E. J. Phares (eds.), *Applications of a social learning theory of personality.* New York: Holt, Rinehart & Winston.

Rotter, J. B., & Rafferty, J. E. (1950). *Manual for the Rotter Incomplete Sentences Blank: College Form.* New York: The Psychological Corporation.

Rotter, J. B., Lah, M., & Rafferty, J. (1992). *Manual—Rotter Incomplete Sentences Blank* (2nd ed.). Orlando, FL: The Psychological Corporation.

Rotter, J. B., Rafferty, J. E., & Schachtitz, E. (1965). Validation of the Rotter Incomplete Sentences Test. In B. I. Murstein (ed.), *Handbook of projective techniques.* New York: Basic Books.

Rounds, J. B. (1985). Vocational Preference Inventory. *The Ninth Mental Measurements Yearbook.* Lincoln, NE: University of Nebraska Press.

Rounds, J. B., Jr., Henly, G. A., Dawis, R. V., Lofquist, L. H., & Weiss, D. J. (1981). *Manual for the Minnesota Importance Questionnaire: A measure of vocational needs and values.* Minneapolis, MN: Vocational Psychology Research, University of Minnesota.

Rousseau, D. M. (1989). Review of the Values Scale, Research Edition. *The Tenth Mental Measurements Yearbook.* Lincoln, NE: University of Nebraska Press.

Ruch, F. L., & Ruch, W. W. (1980). *Employee Aptitude Survey technical report.* Los Angeles: Psychological Services, Inc.

Ruscio, J. (1998). Gatekeeping, compensation, and fallibility. *American Psychologist, 53,* 568-569.

Russell, C. J., Mattson, J., Devlin, S. E., & Atwater, D. (1990). Predictive validity of biodata items generated from retrospective life experience essays. *Journal of Applied Psychology, 75,* 569-580.

Russo, J. (1994). Thurstone's scaling model applied to the assessment of self-reported depressive severity. *Psychological Assessment, 6,* 159-171.

Rust, J., & Lindstrom, A. (1996). Concurrent validity of the WISC-III and Stanford-Binet-IV. *Psychological Reports, 79,* 618-620.

Ryan, A. M., & Sackett, P. R. (1987). Pre-employment honesty testing: Fakability, reactions of test takers, and company image. *Journal of Business and Psychology, 1,* 248-256.

Ryan, J. J., & Lewis, C. V. (1988). Comparison of normal controls and recently detoxified alcoholics on the Wechsler Memory Scale-Revised. *The Clinical Neuropsychologist, 2,* 173-180.

Ryan, J. J., Prifitera, A., & Powers, L. (1983). Scoring reliability on the WAIS-R. *Journal of Consulting and Clinical Psychology, 51,* 460.

Ryan, M. (1985). Review of the Minnesota Clerical Test. *The Ninth Mental Measurements Yearbook* (vol. I). Lincoln, NE: University of Nebraska Press.

Ryan, R. M. (1987). Thematic Apperception Test. In D. J. Keyser & R. C. Sweetland (eds.), *Test critiques compendium.* Kansas City, MO: Test Corporation of America.

Saccuzzo, D. P., & Johnson, N. E. (1995). Traditional psychometric tests and proportionate representation: An intervention and program evaluation study. *Psychological Assessment, 7,* 183-194.

Sackett, P. R., Burris, L. R., & Callahan, C. (1989). Integrity testing for personnel selection: An update. *Personnel Psychology, 42,* 491-529.

Salvia, J., & Ysseldyke, J. (1988). *Assessment in special and remedial education* (4th ed.). Boston: Houghton Mifflin.

Salvia, J., & Ysseldyke, J. (1991). *Assessment* (5th ed.). Boston, MA: Houghton Mifflin.

Samelson, F. (1977). World War I intelligence testing and the development of psychology. *Journal of the History of the Behavioral Sciences, 13,* 274-282.

Sandford, J. A., & Turner, A. (1997). *Intermediate Visual and Auditory Continuous Performance Test* (IVA). Los Angeles: Western Psychological Services.

Sarason, I. G. (1961). Test anxiety, experimental instructions, and verbal learning. *American Psychologist, 16,* 374.

Sarason, I. G. (ed.). (1980). *Test anxiety: Theory, research, and applications.* Hillsdale, NJ: Erlbaum.

Sattler, J. M. (1982). Age effects on Wechsler Adult Intelligence Scale-Revised tests. *Journal of Consulting and Clinical Psychology, 50,* 785-786.

Sattler, J. M. (1988). *Assessment of children* (3rd ed.). San Diego, CA: Jerome M. Sattler, Publisher.

Sattler, J., & Ryan, J. (1999). *Assessment of children, revised and updated third edition, WAIS-III supplement.* San Diego: Jerome M. Sattler, Publisher.

Saxe, L., Dougherty, D., and Cross, T. (1985). The validity of polygraph testing: Scientific analysis and public controversy. *American Psychologist, 40,* 355-366.

Scarr, S. (1981). Testing for children: Assessment and the many determinants of intellectual competence. *American Psychologist, 36,* 1159-1168.

Scarr, S. (1994). Culture-Fair and Culture-Free tests. In R. J. Sternberg (ed.), *Encyclopedia of human intelligence.* New York: Macmillan.

Scarr, S., & Weinberg, R. A. (1976). IQ test performance of black children adopted by white families. *American Psychologist, 31,* 726-739.

Scarr, S., & Weinberg, R. A. (1983). The Minnesota Adoption Studies: Genetic differences and malleability. *Child Development, 54,* 260-267.

Scarr, S., Pakstis, A., Katz, S., & Barker, W. (1977). The absence of a relationship between degree of white ancestry and intellectual skills within a white population. *Human Genetics, 39,* 69-86.

Scarr-Salapatek, S. (1971). Unknowns in the IQ equation. *Science, 174*, 1223-1228.

Schaie, K. W. (1958). Rigidity-flexibility and intelligence: A cross-sectional study of the adult life span from 20-70. *Psychological Monographs, 72*, no. 9 (Whole No. 462).

Schaie, K. W. (1977). Quasi-experimental designs in the psychology of aging. In J. E. Birren & K. W. Schaie (eds.), *Handbook of the psychology of aging*. New York: Van Nostrand Reinhold.

Schaie, K. W. (1978). Review of Senior Apperception Techniques. *The Eighth Mental Measurements Yearbook*. Lincoln, NE: University of Nebraska Press.

Schaie, K. W. (1980). Cognitive development in aging. In L. K. Obler & M. Alpert (eds.), *Language and communication in the elderly*. Lexington, MA: Heath.

Schaie, K. W. (1983). The Seattle Longitudinal Study: A twenty-one year exploration of psychometric intelligence in adulthood. In: K. W. Schaie (ed.), *Longitudinal studies of adult psychological development*. New York: Guilford Press.

Schaie, K. W. (1985). *Manual for the Schaie-Thurstone Adult Mental Abilities Test (STAMAT)*. Palo Alto, CA: Consulting Psychologists Press.

Schaie, K. W. (1994). The course of adult intellectual development. *American Psychologist, 49*, 305-313.

Schaie, K. W., & Willis, S. L. (1986). *Adult development and aging*. Boston, MA: Little, Brown & Co.

Schakel, J. A. (1986). Cognitive assessment of preschool children. *School Psychology Review, 15*, 200-215.

Schear, J. M., & Craft, R. B. (1989). Examination of the concurrent validity of the California Verbal Learning Test. *The Clinical Neuropsychologist, 3*, 162-168.

Scheuneman, J. D. (1987). An argument opposing Jensen on test bias: The psychological aspects. In S. Modgil & C. Modgil (eds.), *Arthur Jensen: Consensus and controversy*. New York: Falmer Press.

Schinka, J., Dye, D., & Curtiss, G. (1997). Correspondence between five-factor and RIASEC models of personality. *Journal of Personality Assessment, 68*, 355-368.

Schlundt, D. G., & Bell, C. (1993). Body Image Testing System: A microcomputer program for assessing body image. *Journal of Psychopathology and Behavioral Assessment, 15*, 267-285.

Schmidt, F. L., Hunter, J. E., McKenzie, R. C., & Muldrow, T. W. (1979). Impact of valid selection procedures on work-force productivity. *Journal of Applied Psychology, 64*, 609-626.

Schmidt, F. L., Hunter, J. E., Outerbridge, A. N., & Trattner, M. H. (1986). The economic impact of job selection methods on size, productivity, and payroll costs of the federal work force: An empirically based demonstration. *Personnel Psychology, 39*, 1-29.

Schmidt, K. L. (1994). Detroit Test of Learning Aptitude-Third Edition. *Journal of Psychoeducational Assessment, 12*, 87-91.

Schmidt, M. (1996). *Rey Auditory-Verbal Learning Test*. Los Angeles: Western Psychological Services.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*, 350-353.

Schmitt, N. (1995). Review of the Differential Aptitude Tests, Fifth Edition. *The Twelfth Mental Measurements Yearbook*. Lincoln, NE: University of Nebraska Press.

Schmitt, N., & Robertson, I. (1990). Personnel selection. *Annual Review of Psychology, 41*, 289-320.

Schoenfeldt, L. F. (1985). Review of Wonderlic Personel Test. *The Ninth Mental Measurements Yearbook* (vol. II). Lincoln, NE: University of Nebraska Press.

Schoenthaler, S. J., Amos, S., Eysenck, H., Peritz, E., & Yudin, J. (1991). Controlled trial of vitamin-mineral supplementation: Effects on intelligence and performance. *Personality and individual Differences, 12*, 351-362.

Schuldberg, D. (1988). The MMPI is less sensitive to the automated testing format than it is to repeated testing: Item and scale effects. *Computers in Human Behavior, 4*, 285-298.

Schwab, L. O. (1979). *The Nebraska assessment for independent living* (Project 93-013). Lincoln, NE: Department of Human Development and the Family, University of Nebraska.

Scott, L. H. (1981). Measuring intelligence with the Goodenough-Harris Drawing Test. *Psychological Bulletin, 89*, 483-505.

Seashore, C. E. (1938). *The psychology of musical talent*. Boston: Silver, Burdett.

Self, P. A., and Horowitz, F. D. (1979). The behavioral assessment of the neonate: An overview. In J. D. Osofsky (ed.), *Handbook of infant development*. New York: Wiley.

Seligman, M. E. P., Abramson, L. Y., Semmel, A., & Von Baeyer, C. (1979). Depressive attributional style. *Journal of Abnormal Psychology, 88*, 242-247.

Sharkey, K. J., & Ritzler, B. A. (1985). Comparing diagnostic validity of the TAT and a new Picture Projective Test. *Journal of Personality Assessment, 49*, 406-412.

Shaughnessy, M., & Moore, J. (1994). The KAIT with developmental students, honor students, and freshmen. *Psychology in the Schools, 31*, 286-287.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Thousand Oaks, CA: Sage.

Shaw, S., Cullen, J., McGuire, J., & Brinckerhoff, L. (1995). Operationalizing a definition of learning

disabilities. *Journal of Learning Disabilities, 28*, 586-597.

Sheldon, W., & Stevens, S. (1942). *The varieties of temperament: A psychology of constitutional differences*. New York: Harper & Brothers.

Shepard, J. W. (1989). Review of the Jackson Vocational Interest Survey. *The Tenth Mental Measurements Yearbook*. Lincoln, NE: University of Nebraska Press.

Shipley, W. C. (1940). A self-administering scale for measuring intellectual impairment and deterioration. *Journal of Psychology, 9*, 371-377.

Shipley, W. C. (1983). *Shipley Institute of Living Scale*. Los Angeles: Western Psychological Services.

Shneidman, E. S. (ed.). (1951). *Thematic test analysis*. New York: Grune & Stratton.

Shucard, D. W., and Horn, J. L. (1972). Evoked cortical potential and measurement of human abilities. *Journal of Comparative and Physiological Psychology, 78*, 59-68.

Shuey, A. M. (1966). *The testing of Negro intelligence* (2nd ed.). New York: Social Science Press.

Shum, D., McFarland, K. A., & Bain, J. D. (1990). Construct validity of eight tests of attention: Comparison of normal and closed head injured samples. *The Clinical Neuropsychologist, 4*, 151-162.

Shurrager, H. C. (1961). *A haptic intelligence scale for adult blind*. Chicago: Illinois Institute of Technology.

Shurrager, H. C., & Shurrager, P. S. (1964). *Manual for the Haptic Intelligence Scale for the Blind*. Chicago: Psychology Research Technology Center, Illinois Institute of Technology.

Siegman, A. W. (1956). The effect of manifest anxiety on a concept formation task, a nondirected learning task, and on timed and untimed intelligence tests. *Journal of Consulting Psychology, 20*, 176-178.

Silverstein, A. B. (1982a). Factor structure of the Wechsler Adult Intelligence Scale-Revised. *Journal of Consulting and Clinical Psychology, 50*, 661-664.

Silverstein, A. B. (1986). Organization and Structure of the Detroit Tests of Learning Aptitude (DTLA-2). *Educational and Psychological Measurement, 46*, 1061-1066.

Silverton, L. (1993). *Adolescent Apperception Cards: manual*. Los Angeles: Western Psychological Services.

Sines, J. O. (1966). Actuarial methods in personality assessment. In B. Maher (ed.), *Progress in experimental personality research* (vol. 3). New York: Academic Press.

Sipps, G. J., Berry, G. W., & Lynch, E. M. (1987). WAIS-R and social intelligence: A test of established assumptions that uses the CPI. *Journal of Clinical Psychology, 43*, 499-504.

Sisson, E. D. (1948). Forced-choice: The new Army rating. *Personnel Psychology, 1*, 365-381.

Sivan, A. B. (1991). *Revised Visual Retention Test: Clinical and experimental applications* (5th ed.). San Antonio, TX: The Psychological Corporation.

Skeels, H. M. (1966). Adult status of children with contrasting early life experiences. *Monographs of the Society for Research in Child Development, 31*(3, Serial No. 105).

Skeels, H. M., & Dye, H. B. (1939). A study of the effects of differential stimulation on mentally retarded children. *Proceedings and Addresses of the American Association on Mental Deficiency, 44*, 114-136.

Skinner, B. F. (1953). *Science and human behavior*. New York: Macmillan.

Skinner, B. F. (1974). *About behaviorism*. New York: Alfred A. Knopf.

Skodak, M., & Skeels, H. M. (1949). A final follow-up study of one hundred adopted children. *Journal of Genetic Psychology, 75*, 85-125.

Slaney, R. B. (1989). Review of the Values Scale, Research Edition. *The Tenth Mental Measurements Yearbook*. Lincoln, NE: University of Nebraska Press.

Smith, A. (1960). Changes in Porteus Maze scores of brain-operated schizophrenics after an eight year interval. *Journal of Mental Science, 106*, 967-978.

Smith, A. (1967). The serial sevens subtraction test. *Archives of Neurology, 17*, 78-80.

Smith, A. (1968). The Symbol Digit Modalities Test: A neuropsychologic test for economic screening of learning and other cerebral disorders. *Learning Disorders, 3*, 83-91.

Smith, A. (1973). *Symbol Digit Modalities Test. Manual*. Los Angeles: Western Psychological Services.

Smith, A., & Kinder, E. (1959). Changes in psychological test performances of brain-operated subjects after eight years. *Science, 129*, 149-150.

Smith, A. M., Gacono, C., & Kaufman, L. (1997). A Rorschach comparison of psychopathic and nonpsychopathic conduct disordered adolescents. *Journal of Clinical Psychology, 53*, 289-300.

Smith, C. R. (1983). *Learning disabilities: The interaction of learner, task, and setting*. Boston, MA: Little, Brown.

Smith, D., Carroll, J. L., & Fuller, G. B. (1988). The relationship between the Millon Clinical Multiaxial Inventory and the MMPI in a private outpatient mental health clinic population. *Journal of Clinical Psychology, 44*, 165-174.

Smith, M. B. (1963). Personal values in the study of lives. In R. W. White (ed.), *The study of lives*. New York: Atherton Press.

Smith, M. L., & Milner, B. (1981). The role of the right hippocampus in the recall of spatial location. *Neuropsychologia, 19*, 781-793.

Smith, M., Delves, T., Lansdown, R., Clayton, B., & Graham, P. (1983). The effects of lead exposure on urban children: The Institute of Child Health/ Southampton Study. *Developmental Medicine and Child Neurology, 25,* 1-54.

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of un-ambiguous anchors for rating scales. *Journal of Applied Psychology, 47,* 149-155.

Smith, T. W., Follick, M. J., Ahern, D. K., & Adams, A. (1986). Cognitive distortion and disability in chronic low back pain. *Cognitive Therapy and Research, 10,* 201-210.

Smither, R. D. (1994). *The psychology of work and human performance* (2nd ed.). New York: HarperCollins.

Snow, J. H. (1992). Review of Luria-Nebraska Neuropsychological Battery: Forms I and II. *The Eleventh Mental Measurements Yearbook.* Lincoln, NE: University of Nebraska Press.

Snyder, D. K., Lachar, D., and Wills, R. M. (1988). Computer-based interpretation of the Marital Satisfaction Inventory: Use in treatment planning. *Journal of Marital and Family Therapy, 14,* 397-409.

Society for Industrial and Organizational Psychology, Inc. (1987). *Principles for the validation and use of personnel selection procedures* (3rd ed.). College Park, MD: Author.

Sokol, R. J., & Clarren, S. K. (1989). Guidelines for use of terminology describing the impact of prenatal alcohol on the offspring. *Alcoholism: Clinical and Experimental Research, 13,* 597-598.

Sontag, L. W., Baker, C., and Nelson, V. (1958). Mental growth and personality development: A longitudinal study. *Monographs of the Society for Research in Child Development, 23* (Whole No. 68).

Sostek, A. J., Buchsbaum, M. S., & Rapoport, J. L. (1980). Effects of amphetamine on vigilance performance in normal and hyperactive children. *Journal of Abnormal Child Psychology, 8,* 491-500.

Sparrow, S. S., Balla, D. A., & Cicchetti, D. V. (1984). *Vineland Adaptive Behavior Scales.* Circle Pines, MN: American Guidance Service.

Spearman, C. (1904). "General intelligence", objectively determined and measured. *American Journal of Psychology, 15,* 201-293.

Spearman, C. (1923). *The nature of "intelligence' and the principles of cognition.* London: Macmillan.

Spearman, C. (1927). *The abilities of man.* New York: Macmillan.

***Special Education Today.* (1985).** ACALD definition of learning disabilities, 2, 1-20.

Sperry, R. W. (1964). The great cerebral commissure. *Scientific American, 210,* 42-52.

Spielberger, C. D. (1984). *State-Trait Anxiety Inventory: A comprehensive bibliography.* Palo Alto, CA: Consulting Psychologists Press.

Spielberger, C. D., and others (1983). *Manual for the State-Trait Anxiety Inventory (STAI, Form Y).* Palo Alto, CA: Consulting Psychologists Press.

Spitz, H. H. (1986). *The raising of intelligence: A selected history of attempts to raise retarded intelligence.* Hillsdale, NJ: Erlbaum.

Spohr, H., & Steinhausen, H. (eds.). (1996). *Alcohol, pregnancy, and the developing child.* Cambridge: Cambridge University Press.

Spranger, E. (1928). *Types of men.* Germany: Max Niemeyer Verlag.

Spreen, O., & Strauss, E. (1998). *A compendium of neuropsychological tests: Administration, norms, and commentary* (2nd ed.). New York: Oxford University Press.

Sprigle, J. E., & Schaefer, L. (1985). Longitudinal evaluation of the effects of two compensatory preschool programs on fourth-through sixth-grade students. *Developmental Psychology, 21,* 702-708.

Springer, S. P., & Deutsch, G. (1985). *Left brain, right brain.* San Francisco: W. H. Freeman.

Stafford-Clark, D. (1971). What Freud really said. New York: Schocken Books.

Stanley, J. C. (1971). Reliability. In R. L. Thorndike (ed.), *Educational measurement.* Washington, DC: American Council on Education.

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist, 6,* 613-629.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69,* 797-811.

Steer, R. A., Beck, A. T., & Brown, G. (1989). Sex differences on the Revised Beck Depression Inventory for outpatients with affective disorders. *Journal of Personality Assessment, 53,* 693-702.

Steers, R. M., & Rhodes, S. R. (1978). Major influences on employee attendance: A process model. *Journal of Applied Psychology, 63,* 391-407.

Stehouwer, R. S. (1987). Beck Depression Inventory. In D. J. Keyser & R. C. Sweetland (eds.), *Test critiques compendium.* Kansas City, MO: Test Corporation of America.

***STEP manual and technical report.* (1980).** Princeton, NJ: Educational Testing Service.

Stephenson, W. (1953). *The study of behavior: Q-technique and its methodology.* Chicago: University of Chicago Press.

Stern, W. L. (1912). Uber die psychologischen Methoden der Intelligenzprufung. American translation by G. M. Whipple (1914). The psychological methods of testing intelligence. *Educational*

Psychology Monographs, no. 13, Baltimore: Warwick & York.

Sternberg, R. J. (1981). Intelligence and nonentrenchment. *Journal of Educational Psychology, 73*, 1-16.

Sternberg, R. J. (1985a). Componential analysis: A recipe. In D. K. Detterman (ed.), *Current topics in human intelligence* (vol. 1). Norwood, NJ: Ablex.

Sternberg, R. J. (1985b). *Beyond IQ: A triarchic theory of human intelligence*. Cambridge: Cambridge University Press.

Sternberg, R. J. (1986). *Intelligence applied: Understanding and increasing your intellectual skills*. San Diego, CA: Harcourt Brace Jovanovich.

Sternberg, R. J. (1994). The triarchic theory of intelligence. In R. J. Sternberg (ed.), *Encyclopedia of human intelligence*. New York: Macmillan.

Sternberg, R. J. (1996). *Successful intelligence*. New York: Simon & Schuster.

Sternberg, R. J. (ed.). (1994). *Encyclopedia of human intelligence* (vol. 1, 2). New York: Macmillan.

Sternberg, R. J., & Detterman, D. K. (eds.). (1986). *What is intelligence? Contemporary viewpoints on its nature and definition*. Norwood, NJ: Ablex.

Sternberg, R. J., & Kaufman, J. C. (1988). Human abilities. *Annual Review of Psychology, 49*, 479-502.

Sternberg, R. J., & Williams, W. (1997). Does the Graduate Record Examination predict meaningful success in the graduate training of psychologists? A case study. *American Psychologist, 52*, 630-641.

Sternberg, R. J., & Zhang, L. (1995). What do we mean by giftedness? A pentagonal implicit theory. *Gifted Child Quarterly, 39*, 88-94.

Sternberg, R. J., Conway, B. E., Ketron, J. L., and Bernstein, M. (1981). People's conceptions of intelligence. *Journal of Personality and Social Psychology, 41*, 37-55.

Stevens, S. S. (1946). On the theory of scales and measurement. *Science, 103*, 677-680.

Stoker, H. (1992). Review of Stanford Achievement Test. *Eleventh Mental Measurements Yearbook*. Lincoln, NE: University of Nebraska Press.

Stokes, G., Mumford, M., & Owens (eds.). (1994). *Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction*. Palo Alto, CA: Consulting Psychologists Press.

Stone, B. J. (1994). Group ability test versus teachers' ratings for predicting achievement. *Psychological Reports, 75*, 1487-1490.

Storandt, M., & Hill, R. D. (1989). Very mild senile dementia of the Alzheimer type: 2. Psychometric test performance. *Archives of Neurology, 46*, 383-386.

Streiner, D. L., Goldberg, J. O., & Miller, H. R. (1993). MCMI-II item weights: Their lack of effectiveness. *Journal of Personality Assessment, 60*, 471-476.

Streissguth, A., Bookstein, F., & Barr, H. (1996). A dose-response study of the enduring effects of prenatal alcohol exposure: birth to 14 years. In H. Spohr & H. Steinhausen (eds.), *Alcohol, pregnancy, and the developing child*. Cambridge: Cambridge University Press.

Streissguth, A., Martin, D., Barr, H., & Sandman, B. (1984). Intrauterine alcohol and nicotine exposure: Attention and reaction time in 4-year-old children. *Developmental Psychology, 20*, 533-541.

Strommen, E. (1988). Confirmatory factor analysis of the Kaufman Assessment Battery for Children: A reevaluation. *Journal of School Psychology, 26*, 13-23.

Strong, E. K. (1927). *Vocational Interest Blank Stanford*, CA: Stanford University Press.

Strong, E. K. (1955). *Vocational interests 18 years after college*. Minneapolis, MN: University of Minnesota Press.

Strub, R. L., & Black, F. W. (1985). *The mental status examination in neurology* (2nd ed.). Philadelphia: F. A. Davis.

Stuss, D. T., Stethem, L. L., Hugenholtz, H., & Richard, M. T. (1989). Traumatic brain injury: A comparison of three clinical tests, and analysis of recovery. *The Clinical Neuropsychologist, 3*, 145-156.

Super, D. E. (1957). *The psychology of careers*. New York: Harper.

Super, D. E. (1968). *Work Values Inventory*. Chicago, IL: Riverside Publishing Company.

Super, D. E. (1970). *Work Values Inventory manual*. Chicago, IL: Riverside Publishing Company.

Super, D. E. (1973). The Work Values Inventory. In D. G. Zytowski (ed.), *Contemporary approaches to interest measurement*. Minneapolis, MN: University of Minnesota Press.

Super, D. E. (1990). A life-span, life-space approach to career development. In D. Brown, L. Brooks, and associates (eds.), *Career choice and development* (2nd ed.). San Francisco, CA: Jossey-Bass.

Super, D. E., & Nevill, D. D. (1986). *The Values Scale, research edition*. Palo Alto, CA: Consulting Psychologists Press.

Super, D. E., Thompson, A. S., Lindeman, R. H., and others (1981). *Career Development Inventory*. Palo Alto, CA: Consulting Psychologists Press.

Sutton, R. S. (1992). Review of the Defining Issues Test. *Eleventh Mental Measurements Yearbook*. Lincoln, NE: The Buros Institute.

Suzuki, L., Meller, P., & Ponterotto, J. (eds.). (1996). *Handbook of multicultural assessment: Clinical, psychological, and educational applications*. Englewood Cliffs, NJ: Prentice-Hall.

Swartz, J. D., Reinehr, R. C., & Holtzman, W. H. (1983). *Holtzman Inkblot Technique, 1956-1982: An annotated bibliography.* Austin, TX: Hogg Foundation for Mental Health.

Sweetland, R. C., & Keyser, D. J. (1987). *Tests: A comprehensive reference for assessments in psychology, education, and business.* Kansas City, MO: Test Corporation of America.

Swensen, C. (1957). Empirical evaluations of human figure drawings. *Psychological Bulletin, 54,* 431-466.

Swensen, C. (1968). Empirical evaluations of human figure drawings: 1957-1966. *Psychological Bulletin, 70,* 20-44.

Swenson, W. M., Rome., H., Pearson, J., and Brannick, T. (1965). A totally automated psychological test: Experience in a medical center. *Journal of the American Medical Association, 191,* 925-927.

Swiercinsky, D. (1978). *Manual for the adult neuropsychological evaluation.* Springfield, IL: Charles C. Thomas.

Swiercinsky, D. (1985). *Testing adults.* Kansas City, MO: Test Corporation of America.

Swinton, S. S., & Powers, D. E. (1983). A study of special preparation on GRE analytical scores and item types. *Journal of Educational Psychology, 75,* 104-115.

Sylvester, R. H. (1913). *The form board test.* Psychological Monographs, 15, no. 65.

Tabachnick, B. G., & Fidell, L. S. (1989). *Using multivariate statistics* (2nd ed.). New York: Harper & Row.

Tallent, N. (1987). Computer-generated psychological reports: A look at the modern psychometric machine. *Journal of Personality Assessment, 51,* 95-108.

Tallent, N. (1993). *Psychological report writing* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Tamkin, A. S., & Scherer, I. W. (1957). What is measured by the "Cannot Say" scale of the group MMPI? *Journal of Consulting Psychology, 21,* 413-417.

Tanner, B. A. (1992). Computer-aided reporting of the results of neuropsychological evaluations of traumatic brain injury. *Computers in Human Behavior, 9,* 51-56.

Tasbihsazan, R., Nettelbeck, T., & Kirby, N. (1997). Increasing Mental Development Index in Australian children: A comparative study of two versions of the Bayley Mental Scale. *Australian Psychologist, 32,* 120-125.

Tasto, D. L., Hickson, R., & Rubin, S. E. (1971). Scaled profile analysis of fear survey schedule factors. *Behavior Therapy, 2,* 543-549.

Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection. *Journal of Applied Psychology, 23,* 565-578.

Teacher, Administrator, and Counselor Manual: Iowa Tests of Educational Development. Forms X-8 and Y-8. 1988.

Teare, J. F., & Thompson, R. W. (1982). Concurrent validity of the Perkins-Binet tests of intelligence for the blind. *Journal of Visual Impairment and Blindness, 76,* 279-280.

Teasdale, G., & Jennett, B. (1974). The Glasgow Coma Scale. *Lancet, 2,* 81.

Tellegen, A., & Ben-Porath, Y. (1992). The new uniform *T* score for the MMPI-2: Rationale, derivation, and appraisal. *Psychological Assessment, 4,* 145-155.

Tendler, A. D. (1930). A preliminary report on a test for emotional insight. *Journal of Applied Psychology, 14,* 123-126.

Teng, S. (1942-43). Chinese influence on the western examination system. *Harvard Journal of Asiatic Studies, 7,* 267-312.

Terkel, S. (1974). *Working: People talk about what they do all day and how they feel about what they do.* New York: Random House.

Terman, L. M. (1916). *The measurement of intelligence.* Boston: Houghton Mifflin.

Terman, L. M., & Oden, M. H. (1959). *Genetic studies of genius: The gifted group at mid-life.* Stanford, CA: Stanford University Press.

Terrell, F., Terrell, S., & Taylor, J. (1981). Effect of race of examiner and cultural mistrust on the WAIS performance of Black students. *Journal of Consulting and Clinical Psychology, 49,* 750-751.

Thomas, R. G. (1985). Review of the Minnesota Clerical Test. *The Ninth Mental Measurements Yearbook* (vol. I). Lincoln, NE: University of Nebraska Press.

Thompson, C. (1949). The Thompson modification of the Thematic Apperception Test. *Journal of Projective Techniques, 13,* 469-478.

Thorndike, E. L. (1912). The permanence of interests and their relation to abilities. *Popular Science Monthly, 81,* 449-456.

Thorndike, E. L. (1918). *The seventeenth yearbook of the National Society for the Study of Education. Pt. II.* Bloomington, IL: Public School Publishing Co.

Thorndike, E. L. (1920a). A constant error in psychological ratings. *Journal of Applied Psychology, 4,* 25-29.

Thorndike, E. L. (1920b). Intelligence and its use. *Harper's Magazine, 140,* 227-235.

Thorndike, E. L. (ed.). (1921). Intelligence and Its Measurement: A Symposium. *Journal of Educational Psychology, 12,* 123-147, 195-216.

Thorndike, R. L., & Hagen, E. P. (1993). *Cognitive Abilities Test, Form 5; Examiner's manual.* Boston: Houghton Mifflin.

Thorndike, R. L., Hagen, E. P., and Sattler, J. M. (1986). *The Stanford-Binet Intelligence Scale: Fourth Edition, Guide for administering and scoring*. Chicago: Riverside Publishing Company.

Thorndike, R. M. (1990). Would the real factors of the Stanford-Binet Fourth Edition please come forward? *Journal of Psychoeducational Assessment, 8*, 412-435.

Thurstone, L. L. (1921). Intelligence. In E. L. Thorndike (ed.), Intelligence and Its Measurement: A Symposium. *Journal of Educational Psychology, 12*, 123-147, 195-216.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. The *Journal of Educational Psychology, 16*, 433-451.

Thurstone, L. L. (1929). Theory of attitude measurement. *Psychological Review, 36*, 222-241.

Thurstone, L. L. (1931). Multiple factor analysis. *Psychological Review, 38*, 406-427.

Thurstone, L. L. (1938). Primary mental abilities. *Psychometric Monographs*, no. 1. Chicago: University of Chicago Press.

Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago: University of Chicago Press.

Thurstone, L. L., & Thurstone, T. (1930). A neurotic inventory. *Journal of Social Psychology, 1*, 3-30.

Thurstone, L. L., & Thurstone, T. (1941). Factorial studies in intelligence. *Psychometric Monographs, No.* 2. Chicago: University of Chicago Press.

Tiffin, J. (1968). *Purdue Pegboard Examiner's Manual*. Chicago: Science Research Associates.

Tinsley, D., Tinsley, H., Boone, S., & Shim-Li, C. (1993). Prediction of scientist-practitioner behavior using personality scores obtained during graduate school. *Journal of Counseling Psychology, 40*, 511-517.

Tombaugh, T., McDowell, I., Kristjansson, B., & Hubley, A. (1996). Mini-Mental State Examination (MMSE) and the Modified MMSE (3MS): A psychometric comparison and normative data. *Psychological Assessment, 8*, 48-59.

Tomkins, S. S. (1947). *The Thematic Apperception Test*. New York: Grune & Stratton.

Tow, P. M. (1955). *Personality changes following frontal leucotomy*. London: Oxford University Press.

Traxler, A. E. (1951). Administering and scoring the objective test. In E. F. Lindquist (ed.), *Educational measurement*. Washington, DC: American Council on Education.

Treffert, D. A. (1989). *Extraordinary people*. London: Bantam Press.

Trevisan, M. S. (1992). Review of GED. *The Eleventh Mental Measurements Yearbook*. Lincoln, NE: University of Nebraska Press.

Trites, R. L. (ed.). (1979). *Hyperactivity in children: Etiology, measurement, and treatment implications*. Baltimore, MD: University Park Press.

Trites, R. L., Blouin, A. G. A., & LaPrade, K. (1982). Factor analysis of the Conners Teacher Rating Scale based on a large normative sample. *Journal of Consulting and Clinical Psychology, 50*, 615-623.

Trull, T. J., Useda, J., Costa, Jr., P., & McCrae, R. (1995). Comparison of the MMPI-2 Personality Psychopathology Five (PSY-5), the NEO-PI, and the NEO-PI-R. *Psychological Assessment, 7*, 508-516.

Trull, T. J., Widiger, T., Useda, J., and others (1998). A structured interview for the assessment of the five-factor model of personality. *Psychological Assessment, 10*, 229-240.

Tryon, R. C. (1957). Reliability and behavior domain validity: Reformulation and historical critique. *Psychological Bulletin, 54*, 229-249.

Tsai, L., & Tsuang, M. (1979). The Mini-Mental State Test and computerized tomography. *American Journal of Psychiatry, 136*, 436-439.

Tulsky, D., Zhu, J., & Ledbetter, M. (1997). *WAIS-III WMS-III technical manual*. San Antonio, TX: The Psychological Corporation.

Tupa, D. J., Wright, M., & Fristad, M. (1997). Confirmatory factor analysis of the WISC-III with child psychiatric inpatients. *Psychological Assessment, 9*, 302-306.

Tweedie, D., & Shroyer, E. H. (eds.). (1982). *The multihandicapped hearing impaired: Identification and instruction*. Washington, DC: Gallaudet College Press.

Tzeng, O. C. S. (1987). Strong-Campbell Interest Inventory. In D. J. Keyser & R. C. Sweetland (eds.). *Test critiques compendium*. Kansas City, MO: Test Corporation of America.

Tzeng, O., Ware, R., & Chen, J. (1989). Measurement and utility of continuous unipolar ratings for the Myers-Briggs Type Indicator. *Journal of Personality Assessment, 53*, 727-738.

Tziner, A. (1985). Review of Values Inventory. *Ninth Mental Measurements Yearbook*. Lincoln, NE: University of Nebraska Press.

Ullman, R. K., Sleator, E., & Sprague, R. (1988). *ADD-H Comprehensive Teacher's Rating Scale* (ACTeRS) (2nd ed.). Champaign, IL: MetriTech, Inc.

Ulrich, L., & Trumbo, D. (1965). The selection interview since 1949. *Psychological Bulletin, 63*, 100-116.

Umberger, F. G. (1987). Peabody Picture Vocabulary Test-Revised. In D. J. Keyser & R. C. Sweetland (eds.), *Test critiques compendium*. Kansas City, MO: Test Corporation of America.

United States Employment Service. (1970). *Manual for the USES General Aptitude Test Battery*. Washington, DC: United States Department of Labor.

Uomoto, J. M., Turner, J. A., & Herron, L. D. (1988). Use of the MMPI and MCMI in predicting

outcome of lumbar laminectomy. *Journal of Clinical Psychology, 44*, 191-197.

U.S. Department of Labor. (1991). *Dictionary of occupational titles* (4th ed.). Washington, DC: Author.

U.S. Department of Labor. (1993). *Selected characteristics of occupations defined in the revised dictionary of occupational titles* (4th ed.). Washington, DC: Author.

U.S. Department of Education. (1977). Definition and criteria for defining students as learning disabled. *Federal Register, 42(250)*, 65083.

Uzgiris, I. C. (1976). Organization of sensorimotor intelligence. In M. Lewis (ed.), *Origins of intelligence: Infancy and early childhood*. New York: Plenum Press.

Uzgiris, I. C. (1983). Organization of sensorimotor intelligence. In M. Lewis (ed.), *Origins of intelligence: Infancy and early childhood* (2nd ed.). New York: Plenum Press.

Uzgiris, I. C., and Hunt, J. M. (1989). *Assessment in infancy: Ordinal scales of psychological development*. Urbana, IL: University of Illinois Press.

Vaillant, G. (1971). Theoretical hierarchy of adaptive ego mechanisms. *Archives of General Psychiatry, 24*, 107-118.

Vaillant, G. (1977). *Adaptation to life: How the best and the brightest came of age*. Boston, MA: Little, Brown.

Vaillant, G. (1992). *Ego mechanisms of defense: A guide for clinicians and researchers*. Washington, DC: American Psychiatric Press.

Vaillant, G., & Vaillant, C. (1990). Natural history of male psychosocial health, XII: A 45-year study of predictors of successful aging at age 65. *American Journal of Psychiatry, 147*, 31-37.

Valdes, G., & Figueroa, R. A. (1994). *Bilingualism and testing: A special case of bias*. Norwood, NJ: Ablex.

Valencia, R. R., & Rankin, R. J. (1988). Evidence of bias in predictive validity on the Kaufman Assessment Battery for Children in samples of Anglo and Mexican American children. *Psychology in the Schools, 25*, 257-263.

Van de Vijver, F., & Harsveld, M. (1994). The incomplete equivalence of the paper-and-pencil and computerized versions of the General Aptitude Test Battery. *Journal of Applied Psychology, 79*, 852-859.

Van Gorp, W. (1992). Review of Luria-Nebraska Neuropsychological Battery: Forms I and II. *The Eleventh Mental Measurements Yearbook*. Lincoln, NE: University of Nebraska Press.

Vance, B., Kitson, D., & Singer, M. (1985). Relationship between the standard scores of PPVT-R and Wide Range Achievement Test. *Journal of Clinical Psychology, 41*, 691-693.

Vance, H., Maddux, C., Fuller, G., & Awadh, A. (1996). A longitudinal comparison of WISC-III and WISC-R scores of special education students. *Psychology in the Schools, 33*, 113-118.

VanderVeer, B., and Schweid, E. (1974). Infant assessment: Stability of mental functioning in young retarded children. *American Journal of Mental Deficiency, 79*, 1-4.

Varma, A., DeNisi, A., & Peters, L. (1996). Interpersonal affect and performance appraisal: A field study. *Personnel Psychology, 49*, 341-360.

Vaughn, S., & Haager, D. (1994). The measurement and assessment of social skills. In G. R. Lyon (ed.), *Frames of reference for the assessment of learning disabilities: New views on measurement issues*. Baltimore, MD: Brookes Publishing.

Verhoeve, M. (1993). *JVIS applications handbook*. Port Huron, MI: Research Psychologists Press.

Vernon, M. C., & Alles, B. F. (1986). Psychoeducational assessment of deaf and hard-of-hearing children and adolescents. In P. J. Lazarus & S. S. Strichart (eds.), *Psychoeducational evaluation of children and adolescents with low-incidence handicaps*. New York: Grune & Stratton.

Vernon, M. C., & Brown, D. W. (1964). A guide to psychological tests and testing procedures in the evaluation of deaf and hard-of-hearing children. *Journal of Speech and Hearing Disorders, 29*, 414-423.

Vernon, P. A. (1994). Reaction time. In R. J. Sternberg (ed.), *Encyclopedia of human intelligence*. New York: Macmillan.

Vernon, P. A., & Mori, M. (1990). Physiological approaches to the assessment of intelligence. In C. R. Reynolds and R. W. Kamphaus (ed.), *Handbook of psychological and educational assessment of children: Intelligence and achievement*. New York: Guilford.

Vernon, P. E. (1950). *The structure of human abilities*. London: Methuen.

Vernon, P. E. (1979). *Intelligence: Heredity and environment*. San Francisco: Freeman.

Viswesvaran, C., Ones, D., & Schmidt, F. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81*, 557-574.

Wade, T. C., & Baker, T. B. (1977). Opinions and use of psychological tests. *American Psychologist, 32*, 874-882.

Wagner, R. (1949). The employment interview: A critical review. *Personnel Psychology, 2*, 17-46.

Wall, R. E., Hinrichsen, G. A., & Pollack, S. (1989). Psychometric characteristics of the Multidimensional Health Locus of Control Scales among psychiatric patients. *Journal of Clinical Psychology, 45*, 94-98.

Wallach, M. A. (1985). Creativity testing and giftedness. In F. D. Horowitz & M. O'Brien (eds.).

The gifted and talented: Developmental perspectives. Washington, DC: American Psychological Association.

Wallbrown, F. H., Carmin, C. N., & Barnett, R. W. (1988). Investigating the construct validity of the Multidimensional Aptitude Battery. *Psychological Reports, 62*, 871-878.

Walls, R. T., Werner, T. J., Bacon, A., & Zane, T. (1799). Behavior checklist. In J. D. Cone & R. P. Hawkins (eds.), *Behavioral assessment: New directions in clinical psychology*. New York: Bruner/Mazel.

Walls, R. T., Zane, T., & Thvedt, J. E. (1979). *The Independent Living Behavior Checklist*. Dunbar, WV: West Virginia Research and Training Center.

Walsh, W. B., & Holland, J. L. (1992). A theory of personality types and work environments. In W. Walsh, R. Price, & K. Craik (eds.), *Person-environment psychology: Models and perspectives*. Hillsdale, NJ: Erlbaum.

Wang, J., & Kaufman, A. (1993). Changes in fluid and crystallized intelligence across the 20- to 90-year age range on the K-BIT. *Journal of Psychoeducational Assessment, 11*, 29-37.

Ward, L. C. (1990). Prediction of verbal, performance, and full scale IQs from seven subtests of the WAIS-R. *Journal of Clinical Psychology, 46*, 436-440.

Watkins, C. E., Jr., Campbell, V., & McGregor, P. (1988). Counseling psychologists' use of the opinions about psychological tests: A contemporary perspective. *The Counseling Psychologist, 16*, 476-486.

Watkins, C., Campbell, V., Nieberding, R., & Hallmark, R. (1995). Contemporary practice of psychological assessment by clinical psychologists. *Professional Psychology: Research and Practice, 26*, 54-60.

Watson, B. U., & Goldgar, D. E. (1985). A note on the use of the Hiskey-Nebraska Test of Learning Aptitude with deaf children. *Language, Speech, and Hearing Services in the Schools, 16*, 53-57.

Watson, C. G., Thomas, D., & Anderson, P. (1992). Do computer-administered Minnesota Multiphasic Personality Inventories underestimate booklet-based scores? *Journal of Clinical Psychology, 48*, 744-748.

Weaver, S. J. (ed.). (1984). *Testing children*. Kansas City, MO: Test Corporation of America.

Wechsler, D. (1932). Analytic use of the Army Alpha examination. *Journal of Applied Psychology, 16*, 254-256.

Wechsler, D. (1939). *The measurement of adult intelligence*. Baltimore, MD: William & Wilkins.

Wechsler, D. (1941). *The measurement of adult intelligence* (2nd ed.). Baltimore, MD: William & Wilkins.

Wechsler, D. (1944). *Measurement of adult intelligence* (3rd ed.). Baltimore: Williams & Wilkins.

Wechsler, D. (1945). A standardized memory scale for clinical use. *Journal of Psychology, 19*, 87-95.

Wechsler, D. (1949). *Manual for the Wechsler Intelligence Scale for Children*. New York: Psychological Corporation.

Wechsler, D. (1952). *The range of human capacities* (2nd ed.). Baltimore: Williams & Wilkins.

Wechsler, D. (1955). *Manual for the Wechsler Adult Intelligence Scale*. New York: Psychological Corporation.

Wechsler, D. (1974). *Manual for the Wechsler Intelligence Scale for Children-Revised*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (1981). *Manual for the Wechsler Adult Intelligence Scale-Revised*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (1989). *Manual for the Wechsler Preschool and Primary Scale of Intelligence-Revised*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (1991). *Manual for the Wechsler Intelligence Scale for Children-III*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (1992). *Wechsler Individual Achievement Test manual*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (1997). *Manual for the Wechsler Adult Intelligence Scale-III*. San Antonio, TX: Psychological Corporation.

Weekes, N. Y. (1994). Sex difference in the brain. In D. W. Zaidel (ed.), *Neuropsychology* (2nd ed.). San Diego, CA: Academic Press.

Weiner, I. B. (1994). The Rorschach Inkblot Method (RIM) is not a test: implications for theory and practice. *Journal of Personality Assessment, 62*, 498-504.

Weiner, I. B. (1996). Some observations on the validity of the Rorschach inkblot method. *Psychological Assessment, 8*, 206-213.

Weiner, I. B., & Kuehnle, K. (1998). Projective assessment of children and adolescents. In A. S. Bellack, & M. Hersen (eds.), *Comprehensive clinical psychology*, (vol. 4). Amsterdam: Elsevier.

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology, 53*, 774-789.

Weiss, D. J. (ed.). (1983). *New horizons in testing: Latent trait theory and computerized adaptive testing*. New York: Academic Press.

Weiss, D. J., & Vale, C. D. (1987). Computerized adaptive testing for measuring abilities and other psychological variables. In J. N. Butcher (ed.), *Computerized psychological assessment: A practitioner's guide*. New York: Basic Books.

Weiss, D. S., Zilberg, N. J., & Genevro, J. L. (1989). Psychometric properties of Loevinger's Sentence

Completion Test in an adult psychiatric outpatient sample. *Journal of Personality Assessment, 53*, 478-486.

Wesman, A. G. (1971). Writing the test item. In R. L. Thorndike (ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.

Westbrook, B. W., & Bane, K. D. (1992). Review of Defining Issues Test. *Eleventh mental measurements yearbook*. Lincoln, NE: University of Nebraska Press.

Wetter, M. W., Baer, R., Berry, D., & Reynolds, S. (1994). The effect of symptom information on faking on the MMPI-2. *Assessment, 2*, 199-208.

Whipple, G. M. (1910). *Manual of mental and physical tests*. Baltimore: Warwick and York.

Whitaker, H. A., & Kahn, H. J. (1994). Brain and language. In D. W. Zaidel (ed.), *Neuropsychology* (2nd ed.). San Diego, CA: Academic Press.

Whitney, D. R., Malizio, A. G., Patience, W. M. (1985). The reliability and validity of the GED Tests. *American Council on Education GED Research Brief*, May, No. 6.

Wholeben, B. E. (1987). Sixteen Personality Factor Questionnaire. In D. J. Keyser & R. C. Sweetland (eds.), *Test critiques compendium*. Kansas City, MO: Test Corporation of America.

Wiebe, D. J., & Smith, T. (1997). Personality and health: Progress and problems in psychosomatics. In R. Hogan, J. Johnson, & S. Briggs (eds.), *Handbook of personality psychology*. San Diego, CA: Academic Press.

Wiedl, K. H., & Carlson, J. S. (1976). The factorial structure of the Raven Coloured Progressive Matrices Test. *Educational and Psychological Measurement, 36*, 409-413.

Wiens, A. N., McMinn, M. R., & Crossen, J. R. (1988). Rey Auditory-Verbal Learning Test: Development of norms for healthy young adults. *The Clinical Neuropsychologist, 2*, 67-87.

Wiesner, W. H., & Cronshaw, S. F. (1988). A meta-analytic investigation of the impact of interview format and degree of structure on the validity of the employment interview. *Journal of Occupational Psychology, 61*, 275-290.

Wiggins, J. (1997). In defense of traits. In R. Hogan, J. Johnson, & S. Briggs (eds.), *Handbook of personality psychology*. San Diego, CA: Academic Press.

Wilkinson, G. S. (1993). *Wide Range Achievement Test-III: Administration manual*. Wilmington, DE: Wide Range, Inc.

Willard, L. S. (1968). A comparison of Culture Fair Test scores with group and individual intelligence test scores of disadvantaged negro children. *Journal of Learning Disabilities, 1*, 30-35.

Williams, J. A., & Williams, J. D. (1985). Kuder General Interest Survey, Form E. In D. J. Keyser & R. C. Sweetland (eds.), *Test critiques* (vol. 2). Kansas City, MO: Test Corporation of America.

Williams, M. (1979). *Brain damage, behaviour, and the mind*. New York: Wiley.

Williamson, L., Campion, J., Malo, S., & others. (1997). Employment interview on trial: Linking interview structure with litigation outcomes. *Journal of Applied Psychology, 82*, 900-912.

Willingham, W. W., Ragosta, M., Bennett, R., and others (1988). *Testing handicapped people*. Boston: Allyn & Bacon.

Willis, S., & Schaie, K. W. (1986). Training the elderly on the ability factors of spatial orientation and inductive reasoning. *Psychology and Aging, 1*, 239-247.

Wills, G. I. (1984). A personality study of musicians working in the popular field. *Personality and Individual Differences, 5*, 359-360.

Willson, V. L. (1989). Review of the Iowa Tests of Basic Skills. *The Tenth Mental Measurements Yearbook*. Lincoln, NE: University of Nebraska Press.

Wilson, B., Alderman, N., Burgess, P., Emslie, H., & Evans, J. (1996). *Behavioral Assessment of the Dysexecutive Syndrome*. Bury St. Edmunds, England: Thame Valley Test Company.

Wilson, M. N. (1994). African Americans. In R. J. Sternberg (ed.), *Encyclopedia of human intelligence*. New York: Macmillan.

Wilson, R. S. (1983). The Louisville Twin Study: Developmental synchronies in behavior. *Child Development, 54*, 298-316.

Wing, H. (1992). Review of the Bennett Mechanical Comprehension Test. *The Eleventh Mental Measurements Yearbook*. Lincoln, NE: University of Nebraska Press.

Winter, D. G., & Stewart, A. J. (1977). Power motive reliability as a function of retest instructions. *Journal of Consulting and Clinical Psychology, 45*, 436-440.

Wirt, R. D., & Broen, W. E., Jr. (1958). *Booklet for the Personality Inventory for Children*. Minneapolis, MN: Authors.

Wirt, R. D., & Lachar, D. (1981). The Personality Inventory for Children: Development and clinical applications. In P McReynolds (ed.), *Advances in psychological assessment* (vol. 5). San Francisco: Jossey-Bass.

Wirt, R. D., Lachar, D., Klinedinst, J. K., & Seat, P. D. (1984). *Multidimensional description of child personality: A manual for the Personality Inventory for Children, Revised 1984*. Los Angeles: Western Psychological Services.

Wirtz, W. (ed.). (1977). *On further examination: Report of the advisory panel on the Scholastic Aptitude Test score decline*. New York: College Entrance Examination Board.

Wisniewski, J. J., & Naglieri, J. A. (1989). Validity of the Draw A Person: A Quantitative Scoring System with the WISC-R. *Journal of Psychoeducational Assessment, 7*, 346-351.

Wissler, C. (1901). The correlation of mental and physical tests. *The Psychological Review*, Monograph Supplement 3(6).

Witkin, H. A. (1949). Perception of body position and of the position of the visual field. *Psychological Monographs, 63*, (Whole No 302).

Wolf, A. W., Schubert, D., Patterson, M., Grande, T., & Pendleton, L. (1990). The use of the MacAndrew Alcoholism Scale in detecting substance abuse and antisocial personality. *Journal of Personality Assessment, 54*, 747-755.

Wolf, T. H. (1973). *Alfred Binet*. Chicago, IL: The University of Illinois Press.

Wolff, K. C., and Gregory, R. J. (1992). The effects of a temporary dysphoric mood upon selected WAIS-R subtests. *Journal of Psychoeducational Assessment, 9*, 340-344.

Wolpe, J. (1958). *Psychotherapy by reciprocal inhibition*. Stanford, CA: Stanford University Press.

Wolpe, J. (1973). *The practice of behavior therapy* (2nd ed.). New York: Pergamon.

Wolpe, J., & Lang, P. J. (1977). *Manual for the Fear Survey Schedule (revised)*. San Diego, CA: Educational and Industrial Testing Service.

Wonderlic, E. F. (1983). *Wonderlic Personnel Test manual*. Northfield, IL: E. F. Wonderlic & Associates.

Wood, J. M., Nezworski, M., & Stejskal, W. (1996). The Comprehensive System for the Rorschach: A critical examination. *Psychological Science, 7*, 3-10.

Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock-Johnson Psycho-Educational Battery-Revised*. Allen, TX: DLM Teaching Resources.

Woodcock, R. W., McGrew, K. S., & Werder, J. K. (1994). *Mini-Battery of Achievement: Examiner's manual*. Chicago, IL: Riverside.

Woodworth, R. S. (1919). Examination of emotional fitness for warfare. *Psychological Bulletin, 16*, 59-60.

Woodworth, R. S. (1951). Autobiography. In C. Murchison (ed.), *History of psychology in autobiography, vol. 2*. New York: Russell and Russell.

Wright, B. D., & Stone, M. (1979). *Best test design*. Chicago, IL: MESA Press.

Wright, J. C., & Mischel, W. (1987). A conditional approach to dispositional constructs: The local predictability of social behavior. *Journal of Personality and Social Psychology, 53*, 1159-1177.

Wright, L. (1988). The Type A behavior pattern and coronary artery disease: Quest for the active ingredients and the elusive mechanism. *American Psychologist, 43*, 2-14.

Wright, L. (1992). Are the physical and TABP risk factors for heart disease unique to CHD? *Journal of Clinical Psychology, 48*, 705-710.

Wright, L., McCurdy, S., & Rogoll, G. (1992). The TUPA Scale: A self-report measure for the Type A subcomponent of time urgency and perpetual activation. *Psychological Assessment, 4*, 352-356.

Wrightsman, L. S., Nietzel, M. T., and Fortune, W. H. (1994). *Psychology and the legal system* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Wulff, D. M. (1996). The psychology of religion: An overview. In E. P. Shafranske (ed.), *Religion and the clinical practice of psychology*. Washington, DC: American Psychological Association.

Wundt, W. (1862). Die Geschwindigkeit des Gedankens. *Gartenlaube*, 263-265.

Yama, M. (1990). The usefulness of human figure drawings as an index of overall adjustment. *Journal of Personality Assessment, 54*, 78-86.

Yarnold, P. R., & Bryant, F. B. (1988). A note on measurement issues in Type A research: Let's not throw out the baby with the bath water. *Journal of Personality Assessment, 52*, 410-419.

Yates, A. (1954). The validity of some psychological tests of brain damage. *Psychological Bulletin, 51*, 359-379.

Yerkes, R. M. (1919). Report of the psychology Committee of the National Research Council. *Psychological Review, 26*, 83-149.

Yerkes, R. M. (ed.) (1921). *Psychological examining in the United States Army. Memoirs of the National Academy of Sciences*, vol. 15.

Zachary, R. A. (1986). *Shipley Institute of Living Scale*. Los Angeles: Western Psychological Services.

Zachary, R A., Crumpton, E., & Spiegel, D. (1985). Estimating WAIS-R IQ from the Shipley Institute of Living Scale. *Journal of Clinical Psychology, 41*, 532-540.

Zajonc, R. B., Markus, H., & Markus, G. B. (1979). The birth order puzzle. *Journal of Personality and Social Psychology, 37*, 1324-1341.

Zavala, A. (1965). Development of the forced-choice rating scale technique. *Psychological Bulletin, 63*, 117-124.

Ziskin, J., and Faust, D. (1988). *Coping with psychiatric and psychological testimony* (vols. 1-3, 4th ed.). Marina Del Ray, CA: Law & Psychology Press.

Zytowski, D. G. (1985). *Kuder Occupational Interest Survey manual supplement*. Chicago: Science Research Associates. ■

Índice de autores

A

Aamodt, M. G., 511
Abbott, R. D., 520
Abel, E. L., 308
Abell, S. C., 271
Abramson, L. Y., 562
Achenbach, T. M., 442, 443
Ackerman, M. J., 465
Ackerman, P. L., 355
Adams, G. A., 487
Adams, K., 666
Adelman, S., 604
Ahern, F. M., 7
Ahrens, J., 248
Aiken, L., 604, 608, 627
Aita, J., 364
Albert, M. L., 377, 381
Albert, S., 598
Alderman, N., 409
Alfonso, V. C., 237
Allen, A., 591
Allen, M. J., 156, 159
Alles, B. F., 55
Alley, P., 291
Allport, G., 30, 556, 564, 589, 590
Allred, E., 308
Altepeter, T. S., 274
Amabile, T., 447
Amelang, M., 591
Ames, L. B., 192
Amos, S., 306
Anastasi, A., 64, 118, 159, 204, 271
Ancoli, S., 658
Anderson, P., 677
Andersson, H. W., 211
Andrew, D. M., 491
Andronikof-Sanglade, A., 599
Anhalt, R., 515
Anthony, J. C., 414
Antoni, M., 639
Appenzeller, T., 370
Appleby, G. S., 493
Archer, D., 655, 658
Archer, P., 422
Argentero, P., 624
Argyle, M., 656
Armitage, S. G., 364
Arnkoff, D., 647
Aronson, J., 699, 700
Aronson, M. K., 401
Arthur, G., 269
Arvey, R. D., 479, 481, 486, 487, 515, 518, 519
Asher, J., 500
Atkinson, L., 232
Atwater, D., 484
Auer, S., 194
Austin, J. T., 508
Awadh, A., 231
Axelrod, B. N., 408
Ayres, R., 203
Azrin, R., 400

B

Bach, P., 411
Backer, T., 629
Bacon, A., 280
Baer, R., 633
Bagby, R. M., 454, 459, 633
Bain, J. D., 398
Baker, C., 209
Baker, E. H., 201
Baker, T. B., 606
Balla, D., 280, 285
Ballard, J., 265
Ballentine, R. D., 482

Ball-Rokeach, S., 556, 558
Baltes, P., 313, 316
Balthazar, E. E., 285
Bancroft, M. J., 200
Bandura, A., 586, 587, 647
Bane, K. D., 563
Banerji, M., 193
Bardos, A., 611
Barker, W., 311
Barkley, R. A., 441, 444
Barnett, R. W., 248
Barr, H., 307
Barrett, P., 47, 308
Batson, C. D., 566 a 568
Bausell, R. B., 146, 154
Bayless, J. D., 408, 409
Bayley, N., 194, 195
Bays, K., 206
Bear, D. M., 377, 378
Bebeau, M., 563
Beck, A. T., 150, 650 a 653
Beck, F., 432
Beck, S., 596
Becker, T. E., 485
Begalla, M. E., 481
Beirne-Smith, M., 186, 264
Belcher, M. J., 490
Bell, C., 654
Bellack, A. S., 646
Bellak, L., 607, 608
Bell-Dolan, D. J., 650
Bellinger, D., 308
Bem, D. J., 584, 591
Benbow, C., 531, 557
Bender, L., 404
Bennett, G. K., 330
Bennett, R., 267
Bennett, T., 392, 398
Ben-Porath, Y., 631, 633
Benson, D. F., 373, 375, 385
Benson, P., 546, 570, 571
Benton, A., 393, 404, 406, 417
Bentz, V. J., 551
Beres, K., 203
Berg, E. A., 408
Berg, R. A., 411
Berger, S. G., 677
Bergin, A., 565
Bergstrom, B., 680
Berk, L. E., 307
Berk, R. A., 39, 88, 89, 110
Bernard, P., 210
Bernreuter, R. G., 29
Bernstein, D. A., 172
Berry, D., 633
Berry, G. W., 220
Bersoff, D. N., 522, 686
Betz, N. E., 549
Bigler, E., 380
Bigley, S. E., 532
Binet, A., 9 a 12, 171
Bishop, C. H., 641
Bishop, E. G., 622
Bishop, J., 400
Black, F. W., 397, 403
Black, J. E., 304
Blaha, J., 248
Blake, R. J., 492
Blakely, T., 432
Blakley, B., 496
Blane, H. T., 657
Blau, A. D., 401
Blau, T. H., 453, 455, 463, 465
Blazer, D. G., 382
Blessed, G., 417
Blin, Dr., 13
Blinkhorn, S., 471
Block, J., 584
Blouin, A., 441
Blum, L., 277
Blumenthal, J. A., 622
Boersma, D., 641
Boes, J., 639
Boese, R., 482
Boggs, D. H., 58
Bolen, L. M., 205, 298
Bolton, B., 547
Bond, L., 64
Bondy, M., 5
Boodoo, G., 312
Bookheimer, S., 386
Boone, S., 546
Borgen, F., 531, 532
Boring, E., 6 a 8
Borkenau, P., 591
Borman, W. C., 480, 486 a 489, 493, 514, 515
Bornstein, M. H., 211
Borstelmann, L. J., 32
Bouchard, T. J., Jr., 302, 312
Bowers, T., 251
Bowler, R., 401
Boyd, T. M., 409
Bracken, B. A., 171, 273
Bradley, P., 151, 152, 638
Bradley, R. H., 423, 425, 427, 428
Bradshaw, J. L., 368, 372, 385
Braithwaite, V., 559
Brannick, M. T., 501
Brannick, T., 665
Brannon, L., 583
Brass, D. J., 501
Brassell, E., 404
Breland, H. M., 297
Brennan, R. L., 94, 95, 103, 108, 110
Breuer, J., 577
Bricklin, B., 467, 468
Briesen, P., 271
Brigham, C. C., 27
Brinckerhoff, L., 429
Brisby, J. A., 424, 425
Briskin, G. J., 404
Broca, P., 383, 384
Brodal, A., 372
Brodsky, S. L., 473
Brody, E. B., 176, 177
Brody, G. H., 176, 177
Broen, W. E., Jr., 641
Bromberg, W., 10

Brooks-Gunn, J., 210, 312
Brown, B. K., 486
Brown, D. W., 55
Brown, F. G., 355
Brown, G., 651, 653
Brown, K., 71
Brown, L., 272
Brown, M. E., 695
Brown, R. T., 293, 297, 309
Browne, C. G., 594
Bruininks, R., 280, 285
Bruyere, S. M., 265
Bryant, F. B., 622
Buchsbaum, M., 396
Buck, J., 31, 611
Bufford, R., 569, 570
Buis, T., 633
Burge, D. A., 650
Burgemeister, B., 277
Burgess, P., 409
Burke, H. R., 257
Burns, A., 184
Buros, O. K., 48
Burris, L. R., 493
Burton, P., 382
Buschke, H., 401
Bushell, D., Jr., 650
Buss, D. M., 589
Butcher, J. N., 631, 633 a 635, 666, 675, 676, 678
Byrne, K. E., 423

C

Cain, L. F., 285
Caldwell, B. M., 423 a 425
Callahan, C., 493
Callahan, L. A., 457
Camara, W. J., 495
Campbell, D. P., 529 a 532, 539
Campbell, D. T., 135
Campbell, J., 220, 343
Campbell, J. P., 125, 148, 162, 508
Campbell, V., 594, 595, 609
Campion, J. E., 480, 487, 499, 500
Campion, M. A., 486
Canfield, A. A., 84
Canivez, G., 232, 241
Cannell, J. J., 359
Cantor, J., 629
Cantrell, J. D., 639
Cantwell, D., 441
Carlson, C. F., 598
Carlson, J. S., 258
Carmin, C. N., 248
Carpenter, C., 400
Carpenter, C. D., 355
Carpenter, M. B., 367, 370
Carretta, T. R., 336
Carroll, D., 471
Carroll, J. L., 603, 639
Carvajal, H., 206
Cascio, W. F., 479, 480, 517, 520
Castellanos, M., 204
Castelli, W. P., 583
Cattell, J. McK., 2, 7, 8, 174
Cattell, R., 178, 234, 254, 316, 576, 587, 588, 623, 624
Cautela, J. B., 649
Ceci, S., 171
Chafetz, M. E., 657
Chaffee, J. W., 4
Chalfant, J. C., 434
Champaigne, L., 432
Chao, G. T., 483 a 485, 489
Chaplin, W. F., 623
Chapman, J. P., 612
Chapman, L. J., 612
Charcot, J. M., 12
Chastain, R. L., 230
Chatman, S., 203
Chelune, G., 241, 408
Chen, J., 621
Chibnall, J., 677
Choca, J. P., 639, 680
Choi, H., 54
Chudy, J., 607
Cicchetti, D. V., 280, 285
Cicerelli, V. G., 46
Cirincione, C., 457
Clark, C. R., 400
Clark, D. A., 651
Clarke, J. M., 377
Clarke-Stewart, A., 36
Clarkin, J. F., 629
Clarren, S. K., 307
Clayton, B., 308
Cleary, T., 294
Cleckley, H., 131
Clemans, W. V., 54, 57, 58
Cleveland, J. N., 508
Clyne-Jackson, S., 693
Cohen, L. D., 313
Cohen, M., 232
Cohen, M. S., 386
Colarelli, S., 487
Colavito, V., 606
Colby, A., 560, 561
Cole, L. J., 291, 293
Cole, N. S., 138
Cole, R. E., 7
Collins, J. M., 495
Colquitt, A. L., 485
Comrey, A. L., 329, 628, 629
Conners, C. K., 441, 442
Conoley, J. C., 2, 48, 243, 348, 651, 653, 666
Conry, R., 228
Constantino, G., 608, 609
Conway, B. E., 172
Conway, J. M., 485, 487
Cook, M., 656
Cook, T. D., 591
Cooley, E. J., 203
Cooper, D., 619, 422
Corcoran, K., 49
Corkin, S., 371
Cornelius, E. T., 481
Cortina, J. M., 105
Cosden, M., 272
Costa, P. T., Jr., 538, 590, 619, 629

Cote, L., 371
Court, J., 100, 257
Coveny, T. E., 278
Cowdery, K. M., 32
Craft, R. B., 401
Crandall, J. E., 133
Crapo, L., 369
Crawford, J. R., 395
Crawford, K., 510
Crawford, M., 496
Crinella, F., 432
Cripe, L., 407
Critchley, M., 406
Cronbach, L. J., 94, 105, 108, 120, 126, 131, 132, 139
Cronshaw, S. F., 487
Cross, T., 472
Crosson, B., 411
Crumpton, E., 251
Cruse, C., 208
Crutcher, M. D., 371
Crystal, H., 401
Culbertson, J., 434
Cullen, J., 429
Cunningham, J. W., 482
Cureton, E. E., 162, 163
Curtiss, G., 539
Cytowic, R. E., 364, 383, 385

D

Dahlstrom, L. E., 631, 635, 636
Dahlstrom, W. G., 631, 635, 636
Damaye, M., 13
Dana, R. H., 605, 639, 697
Darlington, R. B., 641
Das, J. P., 182, 183, 200, 203, 240
Davidshofer, C. O., 484
Davila, M., 380
Davis, C., 278
Davis, G. D., 689
Davis, K., 417
Davis, M. A., 62
Davis, R., 638, 639
Davison, M., 334
Davison, M. L., 563
Dawes, R. M., 669 a 671
Dawis, R. V., 545
Dean, R. S., 385
Deary, I. J., 184
Delaney, H. D., 409, 410
Delis, D. C., 411
Delves, T., 308
Dembroski, T., 583
DeNisi, A., 515
Deri, S., 31
Detterman, D. K., 171, 188
Deutsch, G., 385
DeVito, A. J., 621
Devlin, S. E., 484
Diamond, E. E., 6, 32, 533
Dickens, S., 454
Digman, J., 590
Dikmen, S., 380
DiLalla, L. F., 211
Dillon, H. J., 88
Dillon, R. F., 258
DiMatteo, M. R., 655, 658
Ding, S., 334
Dodds, J., 422
Dodrill, C., 409, 490
Doll, E., 280
Dolliver, R. H., 532
Domino, G., 637
Donahue, M., 570
Donders, J., 241
Donlon, T. F., 149, 339, 340
Donnay, 532
Donnelly, E. F., 137
Dougherty, D., 472
Dowd, E. T., 677
Dowling, L., 271
Drakeley, R. J., 483, 484
Drebing, C., 414
Drummond, R. J., 619
DuBois, P. E., 697
DuBois, P. H., 4, 12, 13, 22, 32, 244, 264
Duker, J., 671, 673
Dumenci, L., 538
Dumont, R., 208, 214
Duncan, G., 312
Dunn, J. A., 274
Dunn, L. M., 274
Dunnette, M. D., 477, 492
Dunst, C. J., 194
Durham, T., 636
Durieux-Smith, A., 210
Dworkin, R., 220
Dye, D., 539
Dye, H. B., 304
Dymond, R. F., 584, 585

E

Eaker, E., 582, 583
Earles, J., 488
Eaton, N., 492
Ebbinghaus, H., 244
Eber, H. W., 577, 624
Eccles, J., 366
Edelbrock, C. S., 444
Edmonds, A., 434
Edwards, A. E., 616, 618, 619
Edwards, J., 184
Efron, R., 383
Eifert, G., 647
Eisdorfer, C., 313
Eisenson, J., 54
Eisenstein, N., 241
Ekman, P., 657, 658
Elaad, E., 472
Elacqua, T., 487
Elfenbein, D., 560
Elliott, C. D., 208, 231
Elliott, S. N., 444
Ellis, A., 611, 650
Ellison, C., 568, 569
Elser, D., 206

Elzey, F. F., 285
Engelhart, C., 241
England, G. W., 545
Ensor, A., 272
Erbaugh, J., 651
Erickson, J., 570
Erlenmeyer-Kimling, L., 220
Eron, L. D., 605
Ertl, J. P., 183
Erwin, F. W., 484
Esquirol, J. E. D., 11
Estes, W. K., 224
Evans, D. A., 381
Evans, J., 382
Evans, R., 248
Ewart, C. K., 586, 587
Exline, R. V., 656, 657
Exner, J. E., Jr., 596, 597, 599
Eyde, L. D., 686, 690, 696
Eysenck, H. J., 171, 183, 184, 306, 588, 626, 627
Eysenck, M. W., 588, 626, 627

F

Fagan, J. F., 210, 211
Fagan, T. K., 171
Faley, R. H., 518
Fancher, R., 8, 12, 14
Fantuzzo, J. W., 215
Faraone, S., 490
Farnham-Diggory, S., 232, 431
Farr, J. L., 507, 510
Faust, D., 472, 473, 669 a 671, 677
Fay, C., 487
Fedio, P., 396
Fehr, B. J., 656, 657
Feinleib, M., 582
Feist, J., 583
Feldman, R. D., 188
Feldt, L. S., 94, 95, 103, 108, 110
Ferris, G., 515
Ferris, S. H., 414
Fidell, L. S., 326, 330
Figueroa, R. A., 701
Filskov, S. B., 671
Finn, S. E., 695, 696
Fischer, J., 49, 570
Fisher, A., 536
Fisher, G. L., 200
Fisher, S., 577
Fisher, T., 432
Fiske, D. W., 135, 486, 587, 590
Fitzgibbons, D., 515
Flanagan, J. C., 508, 510
Fleishman, E. A., 497
Florio, C., 598
Flynn, J. R., 317
Fogle, E., 195
Folstein, M., 414
Folstein, S., 414
Forns-Santacana, M., 205
Forster, A., 432
Fortune, W. H., 452, 691
Foster, C., 359
Foster, R. W., 285
Foster, S. L., 650
Fowler, J., 572, 598
Fowler, P. C., 222
Fowler, R. D., 473, 635, 665
Fox, H. M., 598
Frank, G., 215, 597
Frank, L. K., 593, 594
Franke, W., 5
Frankenburg, W. K., 210, 422
Frankl, V., 564
Franklin, M. E., 649
Frauenheim, J. G., 433
Frechtling, J. A., 88
Frederickson, L. C., 271
Frederiksen, N., 500, 501
French, J. L., 269
Freud, A., 578, 579
Freud, S., 564, 577 a 579
Fridlund, A. J., 658
Friedhandler, B., 579
Friedman, L., 481
Friedman, M., 582, 622
Friedman, S., 36
Friesen, W. V., 657, 658
Friis, S., 648
Fristad, M., 232
Fruchter, B., 84, 111
Fuess, C. M., 338
Fuld, P. A., 401
Fuller, G., 231
Fuller, G. B., 603, 639
Funkenstein, H., 381
Furnham, A., 536

G

Gacono, C., 598
Galaburda, A. M., 383
Gale, A., 184
Galton, F., 6, 7, 76, 174, 594
Garb, H. N., 474, 598, 669
Garbin, M. G., 150, 653
Gardner, H., 171, 185, 383, 448
Gardner, J., 219, 603
Gardner, W., 464
Gass, C., 695
Gasser, M., 334, 508
Gazzaniga, M., 372
Gdowski, C. L., 641, 666, 672
Geary, D. G., 298, 299
Geffen, G., 400
Geffen, L. B., 400
Gelb, S., 20, 21
Genevro, J. L., 601
Genia, V., 566, 572
Georges, M., 562
Gerard, A. B., 467
Gernert, C., 72
Geschwind, N., 383, 384
Gesell, A., 192, 193
Getter, H., 550
Getz, I. R., 563
Gfeller, J., 677

Ghez, C., 368
Ghiselli, E. E., 125, 148, 162, 489, 491
Gibbs, J., 560, 561
Gifford, R., 502
Gilberstadt, H., 671, 673
Gilliam, J. E., 442
Gillis, J., 455
Ginsburg, H., 179
Ginton, A., 472
Glascoe, F. P., 421, 423
Glaser, R., 88
Gleser, G., 94, 126
Glutting, J. J., 215
Goddard, H. H., 20, 21, 697
Goffin, R. D., 502
Gold, J., 400
Goldberg, J., 639
Goldberg, L. R., 473, 588, 589, 671
Goldberg, P., 603
Golden, C., 410, 411
Goldenberg, D. S., 196, 421
Goldfried, M. R., 607
Goldgar, D. E., 272
Golding, S. L., 461
Goldinger, L. A., 493
Goldman, R., 408
Goldsmith, B. Z., 200
Goldstein, D., 195
Goldstein, I. L., 515
Goldstein, K., 364, 382
Gomez-Benito, J., 205
Gonzales, E., 269
Goodenough, F., 7, 12, 23, 31, 269, 270, 697
Goodglass, H., 343, 402
Goodman, D., 485, 487
Goodman, J., 208, 210
Gordon, M., 396
Gordon, R., 466, 467
Gorsuch, R. L., 322
Goslin, D. A., 28, 29
Gothard, S., 455
Gottfredson, G. D., 536, 538
Gottfredson, L., 488, 550
Gough, H., 151, 152, 492, 495, 552, 636, 638, 672
Gould, S., 10, 12, 21
Graber, B., 411
Graham, J. R., 631, 633 a 636
Graham, P., 308
Grande, T., 634
Gray, A., 582
Gray, C. D., 323
Graybill, D., 413, 604
Greenberg, R. P., 577
Greene, J., 153
Greenough, W. T., 304
Gregory, R. J., 42, 60, 72, 97, 102, 155, 215, 228, 291, 308, 330, 344, 365, 385, 404, 406, 407, 682, 694, 695
Greif, E. B., 561
Gresham, F. M., 444, 610
Greve, K., 408
Grisso, T., 466
Gronwall, D., 396
Groth-Marnat, G., 405, 605, 638
Grove, W., 598
Grube, J., 556, 558
Guertin, T., 259
Guidubaldi, J., 196
Guilford, J. P., 84, 111, 180, 447, 587
Guilmette, S., 204
Guion, R. M., 132, 482, 488, 489, 496, 499, 509, 512
Gulliksen, H., 94, 98 a 99, 145
Gunnoe, C., 47, 308
Gutkin, R. B., 298
Guttman, L., 150
Gynther, M. D., 620
Gynther, R. A., 620

H

Haager, D., 434
Haaland, K. Y., 409
Hachinski, V. C., 381
Haddad, F. A., 276
Hagberg, C., 214
Hagen, E., 232, 233, 237, 252, 309
Haier, R. J., 184, 396
Haiken-Vasen, J., 211
Hailey, B. J., 622
Hain, J., 404
Hakel, M. D., 481
Haladyna, T. M., 639
Hall, H. V., 464
Hall, J., 232
Hall, J. A., 655, 658
Hallmark, R., 594, 595, 609
Halpern, D. F., 291
Hambleton, R. K., 156, 159
Hamilton, S. E., 448
Hammeke, T., 410, 411
Hammill, D., 238, 430
Hammontree, S., 606
Hamsher, K., 393, 404, 406, 417
Handler, L., 598
Haney, T., 583
Hankes, E., 664
Hansch, E., 382
Hansen, J. C., 529, 531
Hanson, G. A., 496
Hanson, M., 486, 487
Hargrave, G. E., 138, 492, 636, 638
Harmon, L. W., 87, 531
Harowski, K., 411
Harper, R. G., 250
Harrington, D. M., 447
Harrington, R. G., 49, 246, 249
Harris, D., 269, 270
Harris, M. M., 510
Harrison, D. A., 509, 510
Harsveld, M., 334
Hart, K. E., 622
Hartsough, C., 442
Harvey, R. J., 479, 482, 512, 515
Hathaway, S., 30, 153, 630, 633
Hawkins, K., 490
Hayes, S., 579
Haynes, S. G., 582
Hayslip, B., 314
Hayworth, M., 611

Hazlett, E., 184
Heaton, R. K., 408, 410, 666
Hebb, D., 382
Heckerl, J. R., 433
Hedge, J., 486, 487
Heilbronner, R., 400
Heilbrun, A. B., Jr., 562, 619
Heilbrun, K., 458, 469
Heinemann, A. W., 250
Heinze, M., 466
Helms, J. E., 616
Hendrickson, A. E., 183, 184
Hendrickson, D. E., 189, 184
Herman, D. O., 230, 278
Herriot, P., 483
Herrnstein, R., 311
Herron, L. D., 636
Hersen, M., 646
Hertz, M., 596
Heuvelman, L., 604
Hiatt, D., 138, 492, 636, 638
Hickson, R., 649
Hill, B. K., 280
Hill, E. F., 599
Hill, R. D., 381
Hilliard, A. G., 293
Hilsenroth, M. J., 598
Hinrichsen, G. A., 586
Hiskey, M. S., 272
Hodges, J. R., 414
Hoepfner, R., 447
Hoffart, A., 648
Hoffman, F. J., 433
Hoffman, U., 253
Hogan, J., 493, 496
Hogan, R., 493
Hoge, D. R., 564, 654
Holland, J. L., 534 a 536, 538, 539
Hollon, S. D., 652
Holmes, T., 123
Holtzman, W. H., 599
Holzinger, K. J., 323
Honts, C. R., 472
Honzik, M., 193, 303
Hood, R., 566
Hoover, H. D., 39, 85
Hope, D. A., 648
Horn, J., 178, 184, 234, 316
Horowitz, F. D., 195, 210, 445
Horowitz, I., 463
Hough, L. M., 492
House, A., 413
Houts, A. C., 591
Howell, R. J., 458
Hubley, A., 414
Huffcutt, A. I., 487
Hugenholtz, H., 397
Hughes, J. L., 128
Huhtaniemi, P., 396
Hulin, C. L., 509
Hull, J., 629
Humphreys, L., 171, 294
Hunsberger, B., 566
Hunt, J. McV., 194
Hunter, J. E., 128, 300, 333, 483, 488, 489
Hunter, M., 269
Hunter, R. F., 483, 489
Hurrell, R., 199
Husserl, E., 583
Hutchings, C. A., 62
Hutt, M., 404
Hynd, G. W., 183
Hyne, S., 539

I

Iacono, W., 302, 472
Igbokwe, N. U., 622
Ilg, F. L., 192
Iliff, L., 381
Impara, J. C., 348
Inwald, R., 492
Irvin, J. A., 532
Irwin, P. M., 445
Ivnik, R., 400

J

Jacklin, C., 291
Jackson, D. N., 246, 248, 447, 532, 549, 582, 619, 620, 667
Jackson, R., 294
Jacobson, J., 211
Jacobson, S., 211
Jago, I., 496
Jako, R., 485, 487, 515
James, W., 565
Jarman, R. F., 200, 240
Jasper, H., 378
Jeanneret, P., 480, 481
Jenkins, C. D., 582, 621, 622
Jenkins, J. J., 15
Jenkins, S. J., 200
Jennett, B., 147, 379
Jensen, A. R., 9, 15, 76, 154, 174, 175, 219, 290, 298, 304, 305, 307, 309, 310
Jensen, C. M., 258
Jessell, T. M., 367, 370
Johnsen, S., 272
Johnson, G. H., 675
Johnson, J. H., 665, 675
Johnson, K. A., 274
Johnson, M. B., 436
Johnson, N. E., 259
Johnson, R. C., 7
Johnson, R. G., 249, 251
Johnson, S. T., 339
Johnson, W. G., 461
Johnston, N., 502
Johnston, W. T., 298
Jones, A., 483
Jones, K. L., 307
Jones, R., 541
Joseph, R., 385
Judge, T., 515
Jung, C. G., 594, 621
Jungeblut, A., 340
Jungman, N., 472

K

Kaemmer, B., 635
Kagan, J., 657
Kahn, H. J., 385
Kahn, M. W., 598
Kaiser, H. F., 105
Kalemba, V., 633
Kalesnik, J., 193
Kalichman, S. C., 636
Kamp, J., 492
Kamphaus, R., 201, 420, 444
Kandel, E., 367, 370, 375, 385
Kane, R. L., 393
Kaplan, C., 197
Kaplan, E., 402, 411
Kaplan, S. L., 237
Karnes, F. A., 203
Karr, S., 206
Katz, S., 311
Kaufman, A. S., 71, 144, 145, 200, 201, 203, 205, 214, 215, 218, 228 a 230, 240, 277, 311, 316, 436, 621
Kaufman, J. C., 171, 184, 186
Kaufman, L., 598
Kaufman, N., 144, 145, 200, 201, 203, 240, 436
Kausler, D., 313, 316
Kazdin, A. E., 650
Keenan, J., 404
Keenan, P. A., 641
Keith, T. Z., 206
Keller, R., 511
Kelley, T. L., 28
Kelly, E. L., 486
Kelly, M. P., 400
Kemmerer, B., 666
Kendall, L. M., 512
Kendall, P. C., 652
Kendrick, S., 294
Kennedy, W. A., 309
Kennelly, K. J., 62
Kent, G. H., 594
Kertesz, A., 404
Ketron, J. L., 172
Keyser, D. J., 2, 48, 243
Kierkegaard, S., 583
Kiernan, R. J., 417
Kifer, E., 341
Killian, G. A., 611, 612
Kimbrough, W., 511
Kinder, E., 407
Kinnear, P. R., 323
Kinsbourne, M., 395
Kirby, J. R., 183, 200, 240
Kirby, N., 195
Kirkpatrick, L., 566
Kitson, D., 276
Klebanov, P., 312
Kleiman, L. S., 520
Kleinmuntz, B., 471
Klieger, D. M., 649
Klimoski, R. B., 162, 267, 523
Kline, P., 94, 133, 143, 488, 641
Kline, R. B., 204, 237
Klinedinst, J. K., 640, 641
Klinger, D. E., 665, 666, 676
Klopfer, B., 596, 609
Klopfer, W. G., 32
Klove, H., 410
Kluckhohn, C., 556
Knight, B. C., 201
Knight, R. G., 380
Knigths, R. M., 400
Knill-Jones, R. P., 147
Knobloch, H., 192, 193
Knox, H. A., 22
Koch, W. R., 256
Kohlberg, L., 556, 560, 561, 562
Kohs, S. C., 22
Kolb, B., 364, 370, 375, 377, 378, 388, 413, 433
Kolen, M. J., 39, 85
Koltai, D., 401
Koluchova, J., 304
Konold, T. R., 232
Kornblith, S. J., 653
Koss, E., 371, 380
Kraft, L. M., 200
Kraft, R., 677
Kramer, J. J., 2, 48, 243, 348, 675
Kramer, R., 560
Krieshok, T. S., 246, 249
Kristjansson, B., 414
Kropp, P., 455
Krugman, M., 612
Krumboltz, J. D., 549
Kuder, G. F., 32, 533
Kuehnle, K., 605
Kula, M., 598
Kuncel, N., 343
Kupfermann, I., 370, 385
Kurtines, W., 561
Kush, J., 232

L

La Rue, A., 371, 381
Laatsch, L., 680
Lachar, D., 634, 640, 641, 666, 672
Lacks, P., 404
Lah, M. I., 601
Lai, T. C., 4
Lam, T., 701
Lambert, N. M., 284, 442
Landon, T., 479
Landy, F., 485, 486, 507, 509, 510
Lane, S., 352
Lang, P. J., 649
Lang, W. S., 196, 609
Langston, J. W., 417
Lansdell, H., 137
Lansdown, R., 308
Lanyon, R. I., 598, 677, 678
LaPrade, K., 441
Larson, G. E., 335, 337, 338, 488
Latham, G. P., 487
Laurent, J., 199, 237, 238
Lavin, C., 199
LaVoie, A. L., 608
Law, H., 559

Lawshe, C. H., 120
Layton, W. L., 546
Ledbetter, M., 107, 218, 224, 227, 313, 399, 570
LeDoux, J. E., 372
Lee, C., 276
Lee, M. S., 248
Leeman, G., 550
Lefcourt, H. M., 586
Lefebvre, M. F., 652
Lehman, R., 60, 123
Lehman, R. A., 410
Leiter, R. G., 268
Leland, H., 284
Leli, D. A., 671
Lelos, D., 461
Lemmon, D., 251
LeResche, L., 414
Lerner, J., 431
Levenson, R. L., 206
Levine, S., 285
Levinson, E., 538
Leviton, A., 308
Lewinsohn. P. M., 610, 653, 654
Lewis, C., 105, 536
Lewis, C. V., 400
Lewis, J. F., 87
Lewis, M., 193, 210
Lezak, M., 225, 363, 379, 397, 399, 401 a 403, 406 a 410
Lichtenberg, P., 401
Lichtenstein, R., 193
Lidz, C., 464
Lieberman, M., 560
Likert, R., 150
Lin, Y., 62
Lincoln, A., 621
Lindal, E., 414
Lindeman, R. H., 549
Lindenberger, U., 316
Lindstrom, A., 199
Lindvall, C. M., 76
Lindzey, G., 556, 594
Linn, R., 351, 352
Lipsitt, P. D., 460, 461
Lipsitz, J. D., 220
Lishman, W. A., 308
Little, S., 610
Loevinger, J., 579, 602
Lofquist, L. H., 545
Lohmann, D. F., 258
Longstaff, H. P., 491
Lord, F. M., 94, 158
Lorge, I., 277
Lorr, M., 550
Louttit, C. M., 594
Lovell, M. R., 641
Lowman, R. L., 528, 549 a 551
Lubin, B., 594
Lubinski, D., 531, 557
Lukens, J., 199
Lukin, M. E., 677
Lunz, M., 680
Luria, A., 182, 200, 375, 379
Lykken, D., 302, 471, 472
Lynch, E. M., 220
Lynn, R. L., 306, 312
Lyon, G. R., 431

M

MacAndrew, C., 634
Macciocchio, S. N., 222
Maccoby, E., 291
MacDougall, J., 583
Machover, K., 31, 610
MacMurray, B., 210
MacPhillamy, D. J., 654
Maddi, S. R., 583
Maddux, C., 231
Magoun, H. W., 367
Maher, C., 47, 308
Mahoney, M., 647
Maker, C. J., 448
Malcolm, K. K., 421
Malec, J., 400
Malgady, R. G., 608, 609
Malizia, K., 210
Malizio, A., 355
Malo, S., 487
Malone, A., 192, 193
Maloney, M., 264
Manning, W. H., 297
Mardell-Czudnowski, C. D., 196, 421
Marks, P. A., 584, 671
Markus, G., 339
Markus, H., 339
Markwardt, F. C., 436
Marston, W., 470
Martell, D. A., 458
Martens, B. K., 442
Martin, D., 307
Martin, J. C., 307
Martuza, V. R., 120
Mason, C. F., 251
Massoth, N. A., 206
Masur, D. M., 401
Matarazzo, J., 88, 228, 230, 473, 677, 678, 681, 690
Matavich, M., 238
Matey, C., 269
Mathews, K. A., 583
Mattingley, J. B., 368, 372, 385
Mattis, S., 417
Mattson, J., 484
Maurer, S. D., 487
Maxey, J., 341
Maxwell, J. K., 276
Maxwell, S., 479
Mayeux, R., 385
Mayfield, E. C., 485
McAllister, L. W., 636
McBratnie, B., 382, 407
McCall, R., 208 a 210
McCall, W. A., 38
McCallum, R. S., 203, 238, 273
McCann, J. T., 639
McCarney, S. B., 442
McCarthy, D., 204
McCaulley, M. H., 621
McClean, S. E., 621

McClearn, G. E., 7, 302
McCloy, R., 492
McCord, D., 220
McCormick, E. J., 480 a 482
McCrae, R. R., 538, 563, 590, 619, 629
McCurdy, S., 621
McDermott, P. A., 215
McDonald, A., 459
McDowell, I., 414
McFarland, K., 398
McGarry, A. L., 460, 461
McGlynn, F. D., 648
McGonigle-Gibson, K., 417
McGrath, S. K., 210
McGreevy, M., 457
McGregor, P., 594
McGrew, K. S., 436
McGrew, M. W., 606
McGue, M., 302
McGuire, J., 429
McGurk, F., 293, 294
McHugh, P., 414
McKeachie, W., 62
McKenzie, R., 128
McKey, R. H., 46
McKinley, J. C., 30, 153, 630, 633
McLean, J. E., 215, 218, 230
McMain, S., 463
McMordie, W. R., 397
McMullin, R. E., 647
McNamara, W. J., 128
McNeish, T., 611
McNemar, Q., 43
Mecham, R. C., 480, 481
Medina, A., 702
Medina, N., 2, 348
Meehl, P., 131, 132, 630, 633, 669 a 671
Meichenbaum, D., 647, 650
Meier, S. T., 136
Meier, V. J., 648
Meisels, S. J., 196
Melchert, T. P., 343
Meller, P., 701
Meloy, J. R., 455
Melton, G. B., 344, 452 a 456, 459, 463, 464, 466, 470
Mendelsohn, M., 651
Mensink, D., 203
Menzies, R. J., 463
Mercer, J., 87
Merenda, P. F., 628
Messick, S., 131, 139, 340, 447
Mettelman, B. B., 396
Meyer, G. J., 599
Michael, W. B., 105
Miele, F., 299
Mill, J. S., 12
Miller, D., 665, 666, 676
Miller, H., 639
Miller, L., 268
Miller, L. K., 186
Miller, L. T., 276
Miller, N. M., 586, 587
Miller, T. R., 590
Millman, J., 153
Millon, T., 638, 639
Mills, C., 259
Millsaps, C., 400
Milmoe, S., 657
Milner, B., 371, 378
Minder, C. C., 201
Miner, J. B., 485
Miner, M. G., 485
Minskoff, E., 433
Mirsen, T., 381
Mirsky, A., 396
Mischel, W., 590, 591
Mishkin, M., 370
Mishra, S. P., 71
Mitchell, J. L., 482
Mitchell, J. V., 48, 243, 348
Mitchell, T. W., 162
Mittenberg, W., 400
Moar, K. J., 400
Moberg, D., 568
Mock, J., 651
Mohan, P., 60, 308
Mohs, R., 417
Moore, E., 698
Moore, J., 214
Moore, W. P., 359, 360
Moreland, K. L., 563, 665 a 667, 674, 675
Morgan, C. D., 31, 605
Morgan, S., 398
Morgeson, F. P., 480
Mori, M., 175, 184
Morris, L., 291
Morris, L. W., 62
Morrison, M., 102, 343, 365, 410
Morrison, T., 343
Mortimer, J., 382
Moruzzi, G., 367
Moses, J. A., 411
Moskowitz, D. S., 591
Moss, P. A., 138, 291, 292
Motowidlo, S. J., 487
Motta, R., 610
Mountain, M., 408
Mountjoy, C., 417
Muchinsky, P. M., 499, 500, 510, 560
Mueller, J., 417
Muldrow, T., 128
Mullett, N., 461
Mulvey, E., 464
Mumford, M. D., 480, 483, 484, 497
Mundfrom, D., 427
Mungas, D. M., 620
Munoz, R. F., 653
Munsterberg, H., 470
Murphy, K. R., 273, 337, 484, 489, 508, 512, 515
Murray, C., 310
Murray, H., 31, 605, 606
Myers, I. B., 621

N

Naglieri, J., 183, 203, 269, 271, 276, 610, 611
Nagoshi, C. T., 7
Nanda, H., 94

Narvaez, D., 563
Naugle, R. I., 241
Nauta, W., 379
Naveh-Benjamin, M., 62
Neeb, R., 482
Needleman, H., 47, 308
Neill, D. M., 2, 348
Neisser, U., 312, 317
Nelson, V., 209
Nesselroade, J. R., 302, 313, 624
Nester, M. A., 267
Nettelbeck, T., 195
Netter, B., 598
Nevill, D. D., 547
Nevo, B., 122, 163
Newborg, J., 196
Newcomer, P., 436
Newland, T. E., 278
Newport, K., 404
Nezworski, M., 598
Niaz, U., 414
Nicholson, R. A., 461
Nieberding, R., 694, 609
Nielsen, G., 656
Nietzel, M. T., 452, 691
Nihira, K., 278, 284
Nijenhuis, J., 335
Nilsen, D., 539
NimmoSmith, I., 395
Nisan, M., 562
Nitko, A. J., 352
Nolan, R. F., 201
Novey, M. S., 657
Novick, M. R., 94, 105, 158
Nowinski, J., 550
Nuechterlein, K., 184
Nunnally, J., 97, 143, 156, 159, 193
Nussbaum, D., 459
Nuttall, E. V., 193
Nutting, S., 479

O

Oakland, T. D., 271
O'Bannon, R. M., 493, 494
Obrzut, A., 203
Obrzut, J. E., 203
Odbert, H., 590
Oden, M. H., 448
Oehler-Stinnett, J., 203, 442
O'Hanlon, A. P., 400
O'Hara, M. W., 653
O'Keefe, J., 265
Oldham, G. R., 501
Oles, H. J., 689
Ollendick, T. H., 649
Olsen, B., 648
Olson, H. C., 307
O'Neill, J., 211
Ones, D. S., 343, 493, 495, 510
Opper, S., 179
Ornberg, B., 599
O'Rourke, D. F., 622
Orpen, C., 551
Osborne, R. T., 219
O'Shea, T., 195
Osipow, S. H., 549
OSS Assesment Staff, 40
Oster, H., 658
Oswald, F., 508
Otis, A. S., 24
Outerbridge, A., 129
Overholser, J. C., 639
Owens, W. A., 483, 484
Ownby, R. L., 695

P

Padawer, J., 598
Padilla, A., 702
Paine, C., 594
Pakstis, A., 311
Palmer, S., 267, 523
Paloutzian, R., 568 a 570
Panek, P. E., 314
Panigua, F., 698
Pantle, M., 251
Papez, J. W., 370
Pardaffy, V. A., 512
Parker, K. C., 232
Parmelee, W. M., 603
Pascal, G. R., 404
Pass, R., 482
Paterson, D. G., 15, 22, 510
Patience, W., 355
Patrick, J., 639
Patterson, C., 308
Patterson, M., 634
Patton, J., 186, 264
Pauker, J., 404
Paul, G. L., 650
Paul, J., 102, 365, 409, 410
Paulhus, D., 579
Paulman, R. G., 62
Payne, A. F., 31
Payne, J., 186, 264
Pearson, J., 665
Pearson, K., 7
Pedersen, N. L., 302
Peek, L. A., 466, 467
Pendleton, L., 634
Penfield, W., 378, 379, 382
Pepple, J., 490
Peresie, H., 47, 308
Peritz, E., 306
Perleth, C., 253
Perry, M. V., 472
Pervin, L. A., 587, 589
Peters, L., 515
Petersen, N. S., 39, 85
Peterson, C. A., 639
Peterson, D., 444, 491
Peterson, N. G., 480
Peterson, P., 411
Petrila, J., 452
Pfeiffer, S. J., 276, 417
Phares, E. J., 603
Phelps, L., 272

Phillips, S. E., 267
Piaget, J., 171, 178, 179, 560
Piedmont, R. L., 619, 629
Piersma, H., 639
Pintner, R., 23, 171
Piotrowski, C., 599
Piotrowski, Z. A., 596, 665, 666
Pirozzolo, F., 382
Plake, B., 666, 677
Plant, W., 228
Plaud, J. J., 647
Plomin, R., 211, 302
Pohlmann, J. T., 258
Pollack, S., 586
Pollens, R., 382, 407
Ponterotto, J., 701
Pope, K., 695
Popham, W. J., 88
Porch, B. E., 404
Porteus, S., 22, 189, 407, 697
Potter, E., 492
Powers, D. E., 64
Powers, L., 65
Poythress, N., 452
Prasse, D. P., 274
Prewett, P. M., 238
Price, L., 208
Prifitera, A., 65, 400
Primhoff, E. S., 686
Proctor, T., 54
Prout, H., 276
Purish, A., 411
Pursell, E. D., 486
Pyle, W. H., 23

Q

Quay, H., 444
Quigley, A., 496
Quinones, M., 49

R

Rabinowitz, A., 364
Rafferty, J. E., 601
Ragosta, M., 267
Rahe, R., 123
Raichle, M. E., 386
Rajaratnam, N., 94
Raju, N. S., 352
Ramey, C. T., 306
Ramey, S., 306
Randolph, C., 400
Rankin, R. J., 201
Rapaport, D., 596
Rapoport, J. L., 396
Raskin, D. C., 472
Rausch, H. R., 371
Raven, J., 100, 257
Raven, J. C., 100, 257
Reddon, J. R., 620
Redlich, F., 423
Ree, M. J., 336, 488
Reed, M. T., 273
Reed, R., 47, 308
Reese, H., 313
Reeve, R. R., 269
Reeves, D., 399, 401
Rehm, L. P., 653
Reid, D. B., 400
Reilly, R. R., 483 a 485, 489
Reinehr, R. C., 599
Reisberg, B., 194
Reitan, R., 22, 100, 364, 368, 380, 385, 388, 392, 393, 402, 404, 405
Renzulli, J. S., 445
Reschly, D. J., 280, 284, 298
Rest, J. R., 560 a 563
Rey, A., 400
Rey, G., 404
Reynolds, C. R., 203, 204, 215, 218, 230, 293, 297, 298, 309, 444, 639
Reynolds, S., 633
Rhodes, L., 195, 509
Riccio, C. A., 232
Ricciuti, H. N., 195
Richard, M. T., 397
Richards, L., 458
Richards, P. S., 563, 565
Richmond, J., 423
Ridgeway, V., 395
Rieber, R. W., 6
Rinas, J., 693
Ritzler, B. A., 607
Roberts, R. J., 408
Robertson, G. J., 690
Robertson, H., 461
Robertson, I., 395, 483, 485, 487
Robinson, R., 170
Rock, S. L., 425
Roecker, C., 413
Roesch, R., 461
Rogers, B., 352, 355
Rogers, C. R., 583
Rogers, M., 701
Rogers, P. L., 655, 658
Rogers, R., 454, 458, 633
Rogler, L. H., 608, 609
Rogoff, B., 188
Rogoll, G., 621
Roid, G., 232, 268, 675
Rokeach, M., 555 a 558
Rome, H., 665
Romero, I., 193
Ronan, G. F., 606
Rorschach, H., 30, 31, 577, 595, 596
Rosanoff, A. J., 594
Rose, M. P., 648
Rosen, W. G., 417
Rosenberg, S., 400
Rosenman, R., 582, 622
Rosenthal, R., 655, 657 a 659
Rosenzweig, S., 604
Ross, C., 232
Ross, J., 565
Ross, T., 401
Rosvold, H., 396
Roth, M., 417
Roth, P., 487

Rothermel, R. D., 641
Rothstein, H. R., 484
Rothstein, M., 502
Rotter, J. B., 586, 601
Rounds, J. B., Jr., 535, 546
Rousseau, D. M., 548
Rowland, K., 515
Rubin, S. E., 649
Ruch, F. L., 497, 498
Ruch, W. W., 497, 498
Ruscio, J., 343
Russell, C. J., 484
Russell, J. S., 520
Russell, J. T., 126, 129
Russo, J., 148
Rust, J., 199
Ryan, A. M., 494
Ryan, J., 531
Ryan, J. J., 65, 226, 400
Ryan, M., 491, 492
Ryan, R. M., 605, 606
Ryburn, M., 199, 237, 238

S

Saccuzzo, D., 258, 259
Sackett, P. R., 493, 494
Salvia, J., 111, 193, 197, 265, 437
Samelson, F., 26
Sampson, H., 396
Sanderson, C., 629
Sandford, J. A., 396
Sandman, B., 307
Sandoval, J., 442
Sarason, I., 62, 396
Sattler, J., 54, 56, 61, 113, 195, 206, 226, 229, 231, 233, 237, 257, 265, 278, 309, 435, 695, 697
Sauter, S., 409
Sax, D., 377
Saxe, L., 472
Saxton, J., 417
Scaglione, R., 463
Scarr, S., 196, 260, 305, 306, 311
Schachtitz, E., 603
Schaefer, L., 306
Schafer, E., 183
Schaie, K. W., 134, 177, 178, 314 a 316, 609
Schakel, J. A., 197
Schaubroeck, J., 510
Schauer, S., 253
Schear, J. M., 401
Schell, A., 308
Scherer, I. W., 632
Scheuneman, J. D., 298
Schiebel, D., 629
Schinka, J., 539
Schlundt, D. G., 654
Schmidt, D., 557
Schmidt, F. L., 128, 300, 484, 493, 495, 510
Schmidt, K. L., 240
Schmidt, M., 484
Schmidt, R., 251
Schmitt, N., 105, 333, 483, 485, 487
Schneider, D. L., 495
Schneidman, E. S., 605
Schoendorf, K., 465
Schoenfeldt, L. F., 490
Schoenrade, P., 566
Schoenthaler, S. J., 306
Schreiber, J., 461
Schubert, D., 634
Schuldberg, D., 677
Schulein, M., 411
Schwab, L. O., 284
Schwartz, J., 276, 367, 370
Schweid, E., 195, 210
Sciarrino, J. A., 500
Scott, L. H., 271
Seashore, H., 42, 330
Seat, P. D., 640, 641
Seeman, W., 584, 671
Segal, N., 302
Seguin, E., 11
Seidman, L., 490
Self, P. A., 210
Seligman, M., 652
Semmel, A., 652
Shadish, W. R., Jr., 591
Shaffer, M., 510
Sharkey, K. J., 607
Shaughnessy, M., 214
Shavelson, R. J., 108
Shaw, D., 203
Shaw, E., 464
Shaw, S., 429, 430
Sheldon, K., 433
Sheldon, W., 581
Shepard, J. W., 533
Shepard, K., 422
Shepherd, P. A., 210, 211
Sherbenou, R. J., 272
Sherman, M., 455
Shim-Li, C., 546
Shipley, W. C., 249
Shore, M., 401
Shroyer, E. H., 55
Shucard, D. W., 184
Shuey, A., 309
Shum, D., 398
Shurrager, H. C., 278
Shurrager, P. S., 278
Siegel, B., 184
Siegman, A. W., 63
Silverstein, A., 222, 285
Simon, J. R., 58
Simon, T., 171
Sines, J. O., 671
Singer, M., 276
Sipps, G. J., 220
Sisson, E. D., 512
Sivan, A., 393, 401, 404, 406, 417
Skarlicki, D., 487
Skeels, H. M., 303, 304
Skinner, B. F., 585, 647
Slaney, R. B., 548
Sleator, E., 442
Sliwak, R., 492
Slobogin, C., 452
Smith, A., 397, 407

Smith, A. M., 598
Smith, C. R., 232
Smith, D., 639
Smith, D. W., 307
Smith, G., 400
Smith, H., 410
Smith, J., 569
Smith, L., 570
Smith, M., 308
Smith, M. B., 556
Smith, M. L., 371
Smith, P. C., 512
Smith, T., 583
Smither, R. D., 510, 514, 516
Snow, W., 408
Snyder, D. K., 672
Snyder, J., 204
Sokol, R. J., 307
Sommerville, J., 395
Sontag, L. W., 209
Sostek, A. J., 396
Sparks, C., 484
Sparks, R., 377
Sparrow, S., 280, 285
Spearman, C., 28, 94, 171, 175
Sperry, R., 372
Spiege., D., 251
Spielberger, C., 623
Spitz, H. H., 306
Spohr, H., 308
Sprague, R., 442
Spranger, E., 547, 556
Spreen, O., 393, 401, 406, 409, 417
Sprigle, J. E., 306
Springer, S. P., 385
St. Laurent, C., 598
Stafford-Clark, D., 577
Staley, S., 463
Steadman, H., 457
Steele, C. M., 699, 700
Steer, R. A., 150, 651, 653
Steers, R. M., 509
Stefansson, J., 414
Stefic, E., 550
Stehouwer, R. S., 653
Steinhausen, H., 308
Stejskal, W., 598
Stephenson, W., 584
Stern, G. G., 16
Sternberg, R. J., 170, 171, 184, 186, 214, 343, 446
Stethem, L. L., 397
Stevens, F., 192, 193
Stevens, S. S., 145, 581
Stewart, A. J., 606
Stock, J., 196
Stoddart, C., 400
Stoker, H., 355
Stokes, G., 483, 484
Stone, B. J., 254
Stone, M., 159
Storandt, M., 381
Strand, J., 648
Strauss, E., 401, 409, 410
Streiner, D. L., 639
Streissguth, A., 307
Strommen, E., 203
Strong, E. K., 8, 32, 529, 530
Strub, R. L., 397, 403
Stuck, A., 414
Stuss, D. T., 397
Suess, J. F., 639
Sullivan, M. W., 193, 210
Summers, B., 257
Super, D. E., 547, 548
Suttell, B. J., 404
Sutton, R. S., 563
Suzuki, L., 701
Svinicki, J., 196
Swartz, J. D., 599
Sweetland, R. C., 2, 48, 243
Swensen, C., 610
Swenson, W. M., 665, 666
Swerdlik, M., 199, 237, 238
Swiercinsky, D., 49, 406
Swihart, A., 417
Swineford, F., 323
Swinton, S. S., 64
Sylvester, R. H., 22
Szondi, L., 31, 32
Szucko, J. J., 471

T

Tabachnick, B. G., 326, 330
Talkington, J., 653
Tallent, N., 675, 677, 694
Talley, J., 408
Tamkin, A. S., 632
Tang, C., 184
Tanner, B. A., 675
Tasbihsazan, R., 195
Tasto, D. L., 649
Tatsuoka, M. M., 577, 624
Taulbee, E. S., 609
Taylor, C. B., 586, 587
Taylor, H. C., 126, 129
Taylor, J., 61, 699
Taylor, L., 378
Teachout, M., 488
Teare, J. F., 278
Teasdale, G., 147, 379
Teglasi, H., 606
Tellegen, A., 302, 631, 635
Tendler, A. D., 594
Teng, S., 4
Terkel, S., 541
Terman, L. M., 16, 171, 448
Terrell, F., 61, 699
Terrell, S., 61, 699
Teuber, H., 371
Thoma, S. J., 560, 563
Thomas, D., 677
Thomas, R. G., 492
Thompson, A. S., 549
Thompson, C., 608
Thompson, L. A., 211
Thompson, R. W., 278
Thormahlen, P., 285

Thorndike, E. L., 8, 32, 38, 171, 514, 550
Thorndike, R. L., 233, 252, 309
Thurstone, L. L., 28, 148, 149, 171, 176, 177, 180, 237, 315, 326, 330
Thurstone, T. G., 28, 176, 177
Thvedt, J., 283
Tiffin, J., 410
Tinsley, D., 546
Tinsley, H., 546
Tissot, S., 259
Tivnan, T., 196
Tobin, M., 610
Tombaugh, T., 414
Tomkins, S. S., 605
Tomlinson, B., 417
Tonsager, M. E., 695, 696
Toop, A., 536
Tow, P. M., 407
Trattner, M., 129
Traxler, A. E., 57
Treffert, D. A., 186
Trevisan, M. S., 355
Trites, R. L., 441
Trontel, E. H., 620
Trull, T. J., 590, 629
Trumbo, D., 485
Tryon, R. C., 94, 108
Tsai, L., 414
Tsuang, M., 414, 490
Tucker, G., 241
Tulsky, D., 107, 218, 224, 227, 313, 399
Tupa, D. J., 232
Turner, A., 396
Turner, J. A., 636
Tweedie, D., 55
Tym, E., 417
Tzeng, O., 531, 621
Tziner, A., 560

U

Ulleland, C. N., 307
Ullman, R., 442
Ulmer, D., 582, 622
Ulrich, L., 485
Umberger, F. G., 274
Uomoto, J. M., 636
Useda, J., 590, 629
Uzgiris, I. C., 194

V

Vaillant, C., 579, 581
Vaillant, G., 579, 581
Valdes, G., 701
Vale, C. D., 679, 680
Valencia, R. R., 201
Van de Riet, V., 309
Van de Vijver, F., 334
Van Denburg, E., 639
Van der Flier, H., 335
Van Gorp, W., 413
Vance, B., 231, 276
VanderVeer, B., 195, 210
Vangel, S., 401
Varma, A., 515
Varney, N., 393, 406, 408, 417
Vaughn, S., 434
Ventis, W., 566
Verhoeve, M., 532
Vernon, M. C., 55
Vernon, P. A., 304
Vernon, P. E., 30, 55, 175, 177, 184
Viglione, D., 455, 598
Villanova, P., 508
Viswesvaran, C., 493, 510
Vogt, A., 410
Von Baeyer, C., 652
Von Korff, M., 414
Von Stockert, T., 377
Vosler-Hunter, W., 570

W

Wade, T. C., 606
Wagner, R., 485
Wall, R. E., 586
Wallace, C. S., 304
Wallach, M., 447
Wallbrown, F. H., 248
Wallis, R., 594
Walls, R., 280, 283
Walsh, W. B., 535, 536, 538, 549
Walters, R. H., 586
Wang, J., 240, 316
Ward, C. H., 651
Ward, M., 264
Ward, T., 395
Ware, R., 621
Warner, M. H., 490
Warren, R. L., 411
Watkins, C., 594, 609
Watlington, D. K., 201
Watson, B. U., 272
Watson, C. G., 677
Watz, L., 271
Wayne, K., 251
Weatherman, R., 280
Weaver, S. J., 49
Webb, N. M., 108
Webster, C., 463
Webster, D., 382
Wechsler, D., 54, 60, 71, 76, 171, 197, 199, 215, 225, 226, 229, 230, 248, 313, 399, 436
Wedding, D., 399, 401
Weekes, N. Y., 383
Weinberg, R. A., 305, 306
Weiner, I. B., 596, 598, 605
Weingardner, J., 620
Weinstein, H. P., 629
Weiss, D. J., 159, 666, 679, 680
Weiss, D. S., 601
Welsh, G. S., 631, 636
Werder, J. K., 436
Wernberger, H., 253
Werner, T., 280

Wernicke, C., 384
Wesman, A. G., 154, 294, 330
Westbrook, B. W., 563
Wetter, M. W., 633
Wheelock, J., 398
Whelley, P., 208
Whipple, G. M., 245
Whishaw, I., 364, 370, 375, 377, 388, 413, 433
Whitaker, H. A., 385
White, J. C., 309
Whiteside, L., 427
Whitney, D., 355
Whitney, J., 250
Whitworth, R. H., 298
Wiebe, D. J., 583
Wieber, J., 195
Wiedl, K. H., 258
Wiesner, W. H., 487
Wiggins, J., 587
Wilkinson, G. S., 436
Willard, L. S., 256
Willging, T., 463
Williams, B., 583
Williams, C. L., 633, 634
Williams, J. A., 534
Williams, J. D., 534
Williams, M., 384
Williams, R. E., 508
Williams, T. A., 665, 666, 676
Williams, W., 343
Williamson, L., 487
Willingham, W. W., 267
Willis, S., 134, 315, 316
Willis, W. G., 183
Wills, G. I., 551
Wills, R. M., 672
Willson, V. L., 201, 203, 352
Wilson, B., 409
Wilson, M. N., 312
Wilson, R. S., 208
Wing, H., 491
Winter, D. G., 606
Wirt, R. D., 640, 641
Wirtz, W., 339
Wise, F., 276
Wiske, M., 196
Wisniewski, J. J., 271
Wissler, C., 8, 9
Witkin, H., 47
Wnek, L., 196
Wolf, A. W., 634
Wolf, I., 657
Wolf, T., 12
Wolff, K. C., 696
Wolfson, D., 22, 100, 365, 368, 380, 385, 388, 392, 405
Wolpe, J., 647 a 649
Wonderlic, E. F., 489
Wood, J. M., 598
Woodcock, R., 280, 436
Woodworth, R. S., 8
Worrall, W., 232
Wright, B., 680
Wright, B. D., 159
Wright, J. C., 591
Wright, L., 583, 621, 622
Wright, M., 232
Wrightsman, L. S., 452, 456, 691
Wrightson, P., 396
Wulff, D. M., 565

Y

Yama, M., 610
Yarnold, P. R., 622
Yates, A., 249
Yazzie, C., 276
Yen, W. M., 156, 159
Yerkes, R. M., 23, 26, 27, 215, 244
Yoakum, C., 32
Youngren, M. A., 653
Youniss, R., 550
Yow, B., 195
Ysseldyke, S., 111, 193, 197, 265, 266, 437
Yuen, S., 7

Z

Zachary, R. A., 250, 251
Zajonc, R., 339
Zalewski, C., 599
Zane, T., 280, 283
Zavala, A., 155
Zax, M., 607
Zedeck, S., 125, 148, 162
Zeiss, A. M., 653, 654
Zettel, J., 265
Zhang, L., 446
Zhu, J., 107, 218, 224, 227, 313, 399
Zilberg, N. J., 601
Zilha, E., 381
Zillmer, E., 222
Ziskin, J., 472, 473, 677
Zytowski, D. G., 32
Zyzanski, S., 582, 622

Índice temático

A

AAMR. *Véase Adaptive Behavior Scales*
ABIC. *Véase Adaptive Behavior Inventory for Children*
Ackerman-Schoendorf Scales for Parent Evaluation of Custody (Escalas Ackerman-Schoendorf de Valoración de los Padres para Custodia), 465
Acomodación, 179
ACT. *Véase American College Test*
Adaptive Behavior Inventory for Children (Inventario de Conducta Adaptativa para Niños), 284
Adaptive Behavior Scales (Escalas de Conducta Adaptativa), 284
Afasia, 402
Agnosia visual, 375
Albemarle contra Moody, 520
Alcohol, abuso de, 380
ALPS. *Véase Aphasia Language Performance Scales*
Alzheimer, enfermedad de, 380 a 381, 401
Alzheimer's Disease Assessment Scale (Escala de Evaluación de la Enfermedad de Alzheimer), 417
Amenaza del estereotipo, 699
American College Test (Prueba Universitaria Estadounidense), 340 a 341
Análisis
 de puestos, 478 a 481
 factorial, 28, 136, 322 a 330
 ejes oblicuos en, 330
 ejes ortogonales en, 330
 matriz de correlación en, 323 a 324
 matriz factorial en, 325 a 326
 matriz factorial rotada en, 326
Angiografía cerebral, 387
Ansiedad ante prueba, 62 a 63
Anxious Self-Statements Questionnaire (Cuestionario de Autoafirmaciones de Ansiedad), 652
Aphasia Language Performance Scales (Escalas de Afasia de Desempeño en el Lenguaje), 404
Aphasia Screening Test (Pruebas de Detección de la Afasia), 402 a 403
Aplicación de pruebas, 54 a 60
 grupales, 57
 importancia de la superficie de escritura en, 58
 influencia del examinador, 60 a 61
 sensibilidad hacia las discapacidades en, 55
Apraxia, 403
Área
 de Broca, 383, 384
 de Wernicke, 384
Aritmética (subprueba Wechsler), 220
Armed Services Vocational Aptitude Battery (Batería de Aptitud Vocacional de las Fuerzas Armadas), 335 a 338
Asignación, 45
Asimetría, 76 a 77
Asimilación, 179
ASPECT. *Véase Ackerman-Schoendorf Scales for Parent Evaluation of Custody*
ASVAB. *Véase Armed Services Vocational Aptitude Battery*
ATQ. *Véase Automatic Thoughts Questionnarie*
Ausentismo, 509 a 510
Autoeficacia, 586
Automatic Thoughts Questionnarie (Cuestionario de Pensamientos Automáticos), 652
Automonitoreo, 653

B

Balthazar Scales of Adaptive Behavior (Escalas Balthazar de Conducta Adaptativa), 285
BARS. *Véase Behaviorally Anchored Rating Scale*

BASC. *Véase Behavior Assessment System for Children*
BAT. *Véase Behavioral Avoidance Test*
Batería de Evaluación Kaufman para Niños, 199 a 204
Batería Psicoeducativa Woodcock-Johnson-Revisada, 435, 436
Battelle Developmental Inventory (Inventario de Desarrollo Battelle), 196
Bayley Scales of Infant Development-II (Escalas Bayley de Desarrollo Infantil-II), 194 a 195, 210
BDI. *Véase Beck Depression Inventory*
Beck Depression Inventory (Inventario de Depresión Beck), 651, 653
Behavior Assessment System for Children (Sistema de Evaluación Conductual para Niños), 444
Behavior Observation Scale (Escala de Observación Conductual), 512
Behavioral Avoidance Test (Prueba de Evitación Conductual), 648
Behaviorally Anchored Rating Scale (Escala de Clasificación con Base Conductual), 512
Bernreuter Personality Inventory (Inventario de Personalidad de Bernreuter), 29
BGT. *Véase* Prueba Gestalt de Bender
Bloom Sentence Completion Survey (Examen de Frases Incompletas de Bloom), 602
BOS. *Véase Behavior Observation Scale*
Boston Diagnostic Aphasia Exam (Examen Boston de Diagnóstico de Afasia), 402
BPS. *Véase Bricklin Perceptual Scales*
Bricklin Perceptual Scales (Escalas Perceptuales Bricklin), 467 a 468
Bruininks-Oseretsky Test of Motor Proficiency (Prueba de destreza motora Bruininks-Oseretsky), 285
BSCS. *Véase Bloom Sentence Completion Survey*
Búsqueda independiente, 537 a 539

C

CAI. *Véase Competency Assessment Instrument*
Calificador, sesgo del, 515
Camelot Behavior Checklist (Lista de Verificación Conductual Camelot), 285
Campbell Interest and Skills Survey (Encuesta Campbell de Intereses y Habilidades), 539 a 541
CAPA. *Véase* Evaluación psicológica asistida por computadora
Capacidades mentales primarias, 177
Career Beliefs Inventory (Inventario de Creencias sobre la Profesión), 549
Career Development Inventory (Inventario de Desarrollo Profesional), 549
Career Directions Inventory (Inventario de Orientación Profesional), 549
CAT. *Véase* Pruebas adaptativas computarizadas
CBCL. *Véase Child Behavior Checklist*
CBTI. *Véase* Interpretación de prueba basada en sistemas
CEEB. *Véase* Comité de Examen de Ingreso a la Universidad
Centros de evaluación, 501 a 502
Cerebelo, 367 a 368
Certificación, 45 a 46
CFIT. *Véase Culture Fair Intelligence Test*
Child Behavior Checklist (Lista de Verificación Conductual para Niños), 442, 443 a 445
CI, concepto de, 16
Cinco factores de la personalidad, modelo de los, 588 a 590
CISS. *Véase Campbell Interest and Skills Survey*
Clase social, 424
Clasificación(es), 45
 de expertos, 147
Claves (subprueba Wechsler), 224 a 225
CMMS. *Véase Columbia Mental Maturity Scale*
CMQ. *Véase Common Metric Questionnaire*
Código Penal Modelo, 457
Coeficiente
 alfa, 105
 de confiabilidad, 99 a 100, 101
 de correlación, 100
 de personalidad, 591
CogAT. *Véase Cognitive Abilities Test*
Cognitive Abilities Test (Prueba de Capacidades Cognoscitivas), 251 a 254
Cognitive Errors Questionnaire (Cuestionario de Errores Cognoscitivos), 652
Columbia Mental Maturity Scale (Escala Columbia de Madurez Mental), 276 a 277
Comisión de Oportunidad Equitativa de Empleo, 519
Comité de Examen de Ingreso a la Universidad, 27, 28
Common Metric Questionnaire (Cuestionario Métrico Común), 482
Competencia para estar sujeto a juicio, 458 a 462
Competency Assessment Instrument (Instrumento de Evaluación de la Competencia), 461
Competency Screening Test (Prueba de Detección de la Competencia), 461
Composición de Objetos (subprueba Wechsler), 224
Comprensión (subprueba Wechsler), 220
Comrey Personality Scales (Escalas de Personalidad Comrey), 627 a 629
"Con Trastorno Emocional Grave", 443
Conducta no verbal, 655 a 660
Confiabilidad, 93 a 114
 características inestables y, 109
 coeficiente alfa en, 105
 fórmula Kuder-Richardson 20 en, 106
 intercalificadores, 106
 interpretación de, 110 a 111
 por mitades, 103 a 104
 pruebas referidas a criterio y, 110
 restricción del rango y, 110
 test-retest, 102
 y error de medición, 94 a 99
 estándar, 111 a 114
Confidencialidad, 691
Conmoción cerebral, 397
Conners' Teacher Rating Scales (Escalas Conners de clasificación del Maestro), 441
Consentimiento informado, 692 a 693
Conservación, 179
Constancia de CI, 216
Construcción de pruebas, 143 a 166
Constructo, 131
Contaminación del criterio, 123, 515 a 516
Continuous Performance Test (Prueba de Desempeño Continuo), 395 a 396
Corrección de conjeturas, 59
Corteza cerebral, 372 a 374
CPI. *Véase* Inventario Psicológico de California
CPS. *Véase Comrey Personality Scales*

CPT. *Véase Continuous Performance Test*
Creatividad, 447
Criterio, 122
Cruz Griega, 405
CST. *Véase Competency Screening Test*
CTRS. *Véase Conners' Teacher Rating Scales*
Cuadro de expectativas, 87
Cuerpo calloso, 371 a 372
Cuestionario
de Estilos de Atribución, 652
de Personalidad de Eysenck, 626 a 627
Culture Fair Intelligence Test (Prueba de Inteligencia Culturalmente Justa), 254 a 256
Cuotas (justicia en las pruebas), 301
Curva reactivo-característica, 158 a 159

D

DAB-2. *Véase Diagnostic Achievement Battery-2*
DAS. *Véase Differential Ability Scales*
DAT. *Véase* Test de Aptitudes Diferenciales
Datos biográficos, 483 a 485
Decreto
de 1991 sobre Derechos Civiles, 523
de Educación para Individuos con Discapacidades, 429
sobre los estadounidenses con discapacidades, 266 a 267, 523
Definición(es)
operacionales, 170
real, 170
Defining Issues Test (Prueba de Cuestiones Determinantes), 562 a 563
Demencia, 381, 382
Denver–II, 422 a 423
Desviación estándar, 74
Detección, 45, 420
Detroit Test of Learning Aptitude-4 (Prueba Detroit de Aptitud para el Aprendizaje-4), 238 a 240
Developmental Indicators for the Assessment of Learning-3 (Indicadores del Desarrollo para la Evaluación del Aprendizaje-3), 196
Developmental Indicators for the Assessment of Learning-Revised (Indicadores del Desarrollo para la Evaluación del Aprendizaje-Revisados), 421
DFH. *Véase* Prueba del Dibujo de la Figura Humana
Diagnostic Achievement Battery-2 (Batería de Diagnóstico del Rendimiento-2), 435, 436
Diagnóstico, 46
DIAL-3. *Véase Developmental Indicators for the Assessment of Learning-3*
DIAL-R. *Véase Developmental Indicators for the Assessment of Learning-Revised*
Diferencias
individuales, 696
raciales y CI, 308 a 312
Differential Ability Scales (Escalas de Capacidad Diferencial), 206 a 208
Discapacidades para el aprendizaje, 46, 428 a 435, 437 a 438
Diseño
con Cubos (subprueba Wechsler), 222 a 223
longitudinal, 314
transecuencial, 314
transversal, 313
Dispraxia de construcción, 385
Distribución
de frecuencias, 72 a 73
normal, 74 a 76

E

Early Screening Inventory (Inventario de Detección Temprana), 196
Early Screening Profiles (Perfiles de Detección Temprana), 196
EAS. *Véase Employee Aptitude Survey*
Economía de fichas, 650
Ecuación de regresión, 125
Educational Testing Service, 358
Edwards Personal Preferences Schedule (Inventario Edwards de Preferencia Personal), 616 a 619
EEG.-*Véase* Electroencefalografía
EEOC. *Véase* Comisión de Oportunidad Equitativa de Empleo
Efecto
alcohólico fetal, 307
de Halo, 514 a 515
del Lago Wobegon, 359
EFeQ. *Véase Examinee Feedback Questionnaire*
EIWN-R PR. *Véase* Escala Wechsler de Inteligencia para Niños-Revisada de Puerto Rico
Ejercicio situacional, 500
Ejes
oblicuos, 330
ortogonales, 330
El Yo y los Mecanismos de Defensa, 579
Electroencefalografía, 387
Employee Aptitude Survey (Estudio de Aptitud del Empleado), 497 a 498
Encuesta de Valores de Rokeach, 557 a 558
Entrenamiento, 64
Entrevista, 485 a 487
EPPS. *Véase Edwards Personal Preferences Schedule*
EPQ. *Véase* Cuestionario de Personalidad de Eysenck
Era de los instrumentos de bronce, 5 a 7
Error(es)
administrativos en la calificación, 65
de medición, 94
estándar,
de la diferencia, 114
de medición, 112
del estimado, 125 a 126
sistemático de medición, 97 a 98
Escala(s)
Binet-Simon de 1905, 13 a 15
Binet-Simon de 1908 y 1911, 14 a 15
traducción de Goddard para las, 20
y pruebas con inmigrantes, 20 a 22
C, 84
de calificación gráfica, 510 a 511
de Decisiones Profesionales, 549
de elección forzosa, 512
de estaninas, 83 a 84
de estenes, 83 a 84
de intervalo, 145
de Juicio Moral, 560 a 561
de razón, 146
de Valores, 547 a 548
Guttman, 150
háptica de inteligencia para adultos invidentes, 278

Likert, 150
McCarthy de Aptitudes y Psicomotricidad para Niños, 204 a 206
ordinales, 145
Rogers de Evaluación de la Responsabilidad Criminal, 458, 459
TUPA. *Véase Time Urgency and Perpetual Activation Scale*
Wechsler de Inteligencia,
-Bellevue, 76, 215
para Adultos-III, 226 a 230
para el Nivel Escolar-III, 230 a 232
para los Niveles Preescolar y Primaria-Revisada, 197 a 198
para Niños-Revisada de Puerto Rico, 701
ESI. *Véase Early Screening Inventory*
ESP. *Véase Early Screening Profiles*
Especificidad, 422
Esquemas, 179
Estado mental
al momento de cometer el Delito, 455
examen de, 413
Estándar de cuidado, 693
Estudio de Valores, 30, 556 a 557
EUA contra Georgia Power, 520
Evaluación, 40
conductual, 645 a 660
del Desarrollo de Niños Pequeños, 196
psicológica asistida por computadora, 664
Examen
Army Alfa, 24 a 27
Army Beta, 24 a 27
y diferencias étnicas, 27
de estado mental, 413
sensoperceptual, 393
Examinee Feedback Questionnaire (Cuestionario de Realimentación del Examinado), 163
Experimental Psychology, 8
Expresión facial, 658
Extraversión, 627

F

Factor general de inteligencia (©g'), 175
16 Factores de la Personalidad (16FP), 623 a 625
Fagan's Test of Infant Intelligence (Prueba Fagan de Inteligencia para Lactantes), 211
Faith Maturity Scale (Escala de Madurez de la Fe), 570 a 572
Falsos negativos, 127
Falsos positivos, 127
Fear Survey Schedule (Inventario de Estudio del Temor), 648
Figuras Incompletas (subprueba Wechsler), 221
Finger Localization Test (Prueba de Localización de Dedos), 393, 394
Finger Tapping Test (Prueba de Golpeteo con los Dedos), 102, 409
FMS. *Véase Faith Maturity Scale*
Formación reticular, 367
Fórmula
Kuder-Richardson 20, 106
Spearman-Brown, 104
Frye contra EUA, 453
FTII. *Véase Fagan's Test of Infant Intelligence*
FTT. *Véase Finger Tapping Test*
Fuld Object-Memory Evaluation (Valoración Fuld de Memoria de Objetos), 401
Funciones
cerebrales del hemisferio derecho, 385 a 386
ejecutivas, 406

G

Ganglios basales, 371
GATB. *Véase General Aptitude Test Battery*
GCDAS. *Véase Gesell Child Development Age Scale*
GED. *Véase Test of General Educational Development*
General Aptitude Test Battery (Batería de Prueba de Aptitud General), 333 a 335
General Work Inventory (Inventario de Trabajo General), 482
Genética e inteligencia, 302 a 304
Geriatric Sentence Completion Form (Forma Geriátrica de Frases Incompletas), 602
Gesell Child Development Age Scale (Escala Gesell de Edad del Desarrollo Infantil), 196
Glándula pineal, 369
Graduate Record Exam (Examen de Selección para Posgrado), 342 a 343
Griggs contra Duke Power, 518, 519
Grooved Pegboard Test (Tablero Perforado para Clavijas), 410
Grupo normativo, 70
norma por edad en, 86
normas locales y de subgrupo en, 86 a 87
GWI. *Véase General Work Inventory*

H

Habilidades psicomotrices, 497
Halstead-Reitan Test Battery (Batería de Pruebas Halstead-Reitan), 394
Head Start, 46
Hereditary Genius, 7
Hipocampo, 370
Hipotálamo, 370
Hipótesis
léxica fundamental, 588
proyectiva, 594
Hiskey-Nebraska Test of Learning Aptitude (Prueba Hiskey-Nebraska de Aptitud para el Aprendizaje), 272
Histograma, 73
HIT. *Véase Holtzman Inkblot Technique*
Hoja de Datos Biográficos, 29
Holtzman Inkblot Technique (Técnica Holtzman de Manchas de Tinta), 599 a 600
Home Observation for Measurement of the Environment (Observación en el Hogar para la Medición del Ambiente), 423 a 428
Home Situations Questionnaire (Cuestionario de situaciones hogareñas), 444
Hopelessness Scale (Escala de Desesperanza), 652

I

IDARE. *Véase* Inventario de Ansiedad:Rasgo-Estado
IDEA. *Véase* Decreto de Educación para Individuos con Discapacidades
ILBC. *Véase Independent Living Behavior Checklist*
Imágenes por resonancia magnética, 387

Independent Living Behavior Checklist (Lista de Verificación de la Conducta de Vida Independiente), 283
Índice
de confiabilidad del reactivo, 157
de dificultad del reactivo, 156
de discriminación del reactivo, 159 a 161
de hereditabilidad, 302
de validez del reactivo, 157 a 158
Individualismo
cualificado (justicia en las pruebas), 301
no cualificado (justicia en las pruebas), 300 a 301
Información (subprueba Wechsler), 218
Informes de tiempos en telescopio, 6
Inquiries into Human Faculty and its Development, 7
Inteligencia, 169 a 189
basada en la experiencia, 187
cambios de, con la edad, 312 a 316
cambios generacionales e, 317 a 320
componencial, 186, 187
contextual, 188
contribuciones genéticas a la, 302 a 304
cristalizada, 178
definiciones de, 170 a 174
diferencias raciales en, 308 a 312
efectos sobre la,
ambientales, 304 a 308
teratogénicos, 307 a 308
factores en, 175
fluida, 178
modelo de la Estructura Intelectual en, 180 a 182
potencial evocado promedio e, 184
procesamiento simultáneo y sucesivo en, 182 a 183
social, 550
teoría de,
múltiple, 185 a 186
triárquica, 186 a 189
Interacción visual, 656 a 657
International Health Exhibition de 1884, 7
Interpretación
de los Sueños, 577
de prueba basada en sistemas, 665
Introversión, 627
Inventario(s)
de Ansiedad:Rasgo-Estado, 622 a 623
de Personalidad de Thurstone, 29
de Valores, 559 a 560
Gesell del Desarrollo, 192 a 193
Multiaxial Clínico de Millon-II, 638 a 639
Multifásico de la Personalidad de Minnesota, 30, 630, 631, 667, 669
Multifásico de la Personalidad de Minnesota-2, 630 a 638
NEO de Personalidad-Revisado, 629
Psicológico de California, 494, 636 a 638
Tipológico de Tipos de Myers-Briggs, 621
Iowa Tests of Basic Skills (Pruebas Iowa de Habilidades Básicas), 351 a 352
Iowa Tests of Educational Development (Pruebas Iowa de Desarrollo Educativo), 352 a 354
IRM. *Véase* Imágenes por resonancia magnética
ITBS. *Véase Iowa Tests of Basic Skills*
ITED. *Véase Iowa Tests of Educational Development*

J

Jackson Vocational Interest Survey (Encuesta Jackson de Interés Vocacional), 532 a 533
JAS. *Véase Jenkins Activity Survey*
Jenkins Activity Survey (Encuesta Jenkins de Actividad), 621, 622
Juicio
actuarial, 669
clínico, 669
Justicia de prueba
cuotas y, 301
e individualismo,
cualificado, 301
no cualificado, 300 a 301
JVIS. *Véase Jackson Vocational Interest Survey*

K

K-ABC. *Véase* Batería de Evaluación Kaufman para Niños
KAIT. *Véase Kaufman Adolescent and Adult Intelligence Test*
Kaufman Adolescent and Adult Intelligence Test (Prueba Kaufman de inteligencia para adolescentes y adultos), 214
Kaufman Test of Educational Achievement (Prueba Kaufman de Rendimiento Educativo), 435 a 437
K-BIT. *Véase* Test Breve de Inteligencia de Kaufman
KGIS. *Véase Kuder General Interest Survey*
Kohs Block Design Test (Prueba de Diseño con Cubos de Kohs), 22
K-TEA. *Véase Kaufman Test of Educational Achievement*
Kuder General Interest Survey (Encuesta de Intereses Generales de Kuder), 32, 533 a 534

L

Laberintos (subprueba Wechsler), 225
Law School Admission Test (Prueba de Admisión a la Facultad de Leyes), 344
Leiter International Performance Scale-Revised (Escala Internacional Leiter de Ejecución-Revisada), 267 a 269
Lenguaje como función del hemisferio izquierdo, 383 a 385
Lesión(es)
cerebral traumática, 5, 380
personales, 469 a 470
Ley Pública 93-112, 264
Ley Pública 94-142, 265 a 266
Ley Pública 99-457, 266
Line Tracing Task (Tarea de Trazado de Líneas), 410
Lista de verificación de incidentes críticos, 510
LNNB. *Véase Luria-Nebraska Neuropsychological Battery*
Lóbulo(s)
frontales, 378
límbico, 370 a 371
occipitales, 375
parietales, 375 a 377
temporales, 377 a 378
Locus de control, 586
LSAT. *Véase Law School Admission Test*
Luria-Nebraska Neuropsychological Battery (Diagnóstico Neuropsicológico Luria-Nebraska), 410 a 413

M

MAB. *Véase Multidimensional Aptitude Battery*
Mal de Parkinson, 371, 381 a 382
Manejo de contingencias, 650
Manual del usuario, 165

Manual técnico, 165
MAT. *Véase* Prueba Metropolitan de Rendimiento
Matrices (subprueba Wechsler), 223
Matrices Progresivas de Raven, 256 a 259
Matriz
de correlación, 323 a 324
factorial, 325 a 326
multirrasgo-multimétodo, 135
MBA. *Véase Mini-Battery of Achievement*
MBTI. *Véase* Inventario Tipológico de Tipos de Myers-Briggs
MCAT. *Véase Medical College Admission Test*
MCMI-II. *Véase* Inventario Multiaxial Clínico de Millon-II
MCT. *Véase Minnesota Clerical Test*
Measures for Clinical Practice: A Sourcebook, 49
Mecanismos de defensa, 579 a 580
Media, 73
Mediana, 73
Medical College Admission Test (Prueba de Admisión a la Facultad de Medicina), 343 a 344
Medidas de atención y concentración, 441 a 442
Médula ósea, 366, 368
Mejores intereses del cliente, 690 a 691
Mental Measurements Yearbook, 48
Mental Tests and Measurements, 2, 8
Mesencéfalo, 368
Metencéfalo, 366, 368
Método
de concordancia empírica, 150 a 151
de escalamiento, 145 a 152
absoluto, 149 a 150
racional, 151 a 152
de intervalos de aparición balanceada, 148 a 149
Miner Sentence Completion Scale (Escala Miner de oraciones incompletas), 602
Mini-Battery of Achievement (Minibatería de Rendimiento), 435, 436
Mini-Mental State Examination (Examen de Estado Mini-mental), 414 a 417
Minnesota Clerical Test (Prueba Minnesota de Trabajos de Oficina), 491 a 492
Minnesota Importance Questionnaire (Cuestionario Minnesota de Importancia), 544 a 546
Minnesota Rate of Manipulation Test (Prueba Minnesota de Tasa de Manipulación), 498
Minorías culturales y lingüísticas, 697 a 702
MIQ. *Véase Minnesota Importance Questionnaire*
MMPI. *Véase* Inventario Multifásico de la Personalidad de Minnesota
MMPI-2. *Véase* Inventario Multifásico de la Personalidad de Minnesota-2
MMSE. *Véase Mini-Mental State Examination*
Moda, 73
Modelo RIASEC, 535
MSCA. *Véase* Escalas McCarthy de Aptitudes y Psicomotricidad para Niños
MSO. *Véase* Estado mental al momento de cometer el Delito
Muestra
de estandarización, 39
de trabajo, 499
Muestreo aleatorio, 85
estratificado, 85
Multidimensional Aptitude Battery (Batería Multidimensional de Aptitud), 246 a 249
Multilingual Aphasia Examination (Examen Multilingüe de la Afasia), 404
Multimedia, 674
Myart contra Motorola, 518 a 519

N

National Teacher Examination (Examen Nacional para Maestros), 357
NEO-PI-R. *Véase* Inventario NEO de Personalidad-Revisado
Neuropsicología, 364
Niveles de medición, 145 a 146
No culpable a causa de insania, 455
Normas éticas y profesionales, 686 a 696
NTE. *Véase National Teacher Examination*

O

OAI. *Véase Occupation Analysis Inventory*
Occupation Analysis Inventory (Inventario de Análisis Ocupacional), 482
Ordenación de Dibujos (subprueba Wechsler), 222
Ordinal Scales of Psychological Development (Escalas Ordinales de Desarrollo Psicológico), 194

P

Paced Auditory Serial Addition Task (Tarea Rítmica Auditiva de Suma en Serie), 396 a 397
Paradoja proyectiva, 594, 612 a 613
Paralingüística, 657
Parent-Child Relationship Inventory (Inventario de Relaciones Padres-Hijo), 467
Pares craneales, 369
PASAT. *Véase Paced Auditory Serial Addition Task*
Patrón conductual Tipo A con propensión a la cardiopatía, 582 a 583, 622
Pautas Uniformes para la Selección de Personal, 520 a 522
Peabody Individual Achievement Test-Revised (Prueba Peabody de Ejecución Individual-Revisada), 435, 436
Peabody Picture Vocabulary Test-III (Prueba Peabody de Vocabulario con Figuras-III), 274 a 276
Percentil, 77 a 78
Perception-of-Relationship Test (Prueba de Percepción de Relaciones), 468
Personalidad, 576 a 577
Personality Inventory for Children (Inventario de Personalidad para Niños), 639 a 641
Personality Projection in the Drawing of the Human Figure, 31
Personality Research Form (Forma de Investigación de la Personalidad), 618 a 621
PES. *Véase Pleasant Events Schedule*
PIAT-R. *Véase Peabody Individual Achievement Test-Revised*
PIC. *Véase Personality Inventory for Children*
PICA. *Véase Porch Index of Communicative Ability*
Picture Projective Test (Prueba de Láminas Proyectivas), 606 a 607
Placer, principio del, 578
Pleasant Events Schedule (Inventario de Acontecimientos Placenteros), 654
PMPQ. *Véase Professional and Managerial Position Questionnaire*
Polígono de frecuencias, 73
PONS. *Véase Profile of Nonverbal Sensitivity*
Porch Index of Communicative Ability (Índice Porch de Habilidad de Comunicación), 404
PPVT-III. *Véase Peabody Picture Vocabulary Test-III*
PRF. *Véase Personality Research Form*

Problema del criterio, 508
Producción divergente, 182, 447
Professional and Managerial Position Questionnaire (Cuestionario de Puestos Profesionales y Gerenciales), 482
Profile of Nonverbal Sensitivity (Perfil de Sensibilidad No Verbal), 658 a 660
Programa de exámenes universitarios, 356
Prosencéfalo, 369
Prueba(s)
adaptativas computarizadas, 679
Bennett de Comprensión Mecánica, 490 a 491
Benton de Retención Visual, 401
culturalmente justas, 259 a 260
de aprovechamiento, 42
de aptitud, 28 a 29, 42
de bandeja de entrada, 500 a 501
de creatividad, 42
de Fichas, 404
de frases incompletas, 601 a 604
de integridad, 494 a 496
de inteligencia, 41
No Verbal-3, 272 a 273
validez predictiva de, con lactantes y preescolares, 208 a 210
de Interpretación de Formas de Rorschach, 30, 31, 595 a 599
de la habitación inclinada y la silla inclinada, 47
de opción múltiple, 154 a 155
de polígrafo, 470 a 472
definición de, 36 a 39
del Dibujo de la Figura Humana, 609 a 611
efectos del entrenamiento sobre las, 64 a 65
futuro de la aplicación de, 680 a 682
Gestalt de Bender, 404
grupales, 23 a 27, 40, 243 a 261, 347 a 360
de rendimiento, 347 a 360
individuales, 40
de rendimiento, 348
ipsativa, 618
Metropolitan de Rendimiento, 352
no verbales a principios del siglo XX, 22
normas o estándares en, 38
procedimiento estandarizado en, 37
proyectivas, 593 a 613
psicológicas, historia de las, 1 a 33
quién puede obtener las, 37
referidas a criterio, 88 a 89, 348 a 349
referidas a la norma, 348 a 349
contra referidas al criterio, 39, 348 a 349
Wonderlic de Personal, 489 a 490
Psicología forense, 451 a 474
Psicómetra, 36
Puntuación(es)
estándar, 78 a 82
normalizadas, 82 a 83
naturales, 70 a 72
T, 80 a 82
verdadera, 38, 94

R

Rapport, 60
Rasgo(s), 587
fuente, 588
superficiales, 587
R-CRAS. *Véase* Escalas Rogers de Evaluación de la Responsabilidad Criminal
Realidad
principio de, 578
virtual, 675
Registro de Preferencias de Kuder, 32
Regla de Durham, 456
Regla M'Naughten, 456
Reitan-Indiana Aphasia Screening Test (Prueba Reitan-Indiana de Detección de la Afasia), 403
Relaciones cerebro-conducta, 364, 392 a 393
Religión como búsqueda, 566 a 568
Religiosidad
extrínseca, 565
intrínseca, 565
Retención de dígitos (subprueba Wechsler), 218 a 219
Retraso mental, 10 a 11, 278 a 280
Revised Behavior Problem Checklist (Lista de Verificación de Problemas Conductuales Revisada), 444
Rey Auditory Verbal Learning Test (Prueba Rey de Aprendizaje Auditivo Verbal), 400
RISB. *Véase Rotter Incomplete Sentences Blank*
Rosenzweig Picture Frustration Study (Estudio Rosenzweig de Frustración por Ilustraciones), 603 a 604
Rotación
hacia estructura simple, 326
hacia variedad positiva, 326
Rotter Incomplete Sentences Blank (Formulario de Frases Incompletas de Rotter), 601 a 603
RPM. *Véase* Matrices Progresivas de Raven

S

SAchT. *Véase Stanford Achievement Test*
SAT. *Véase Scholastic Assessment Test*
SB:CE. *Véase* Stanford-Binet: Cuarta Edición
Scales of Independent Behavior-Revised (Escalas de Conducta Independiente-Revisadas), 280 a 282
Scholastic Assessment Test (Pruebas de Evaluación Escolar), 338 a 340
Science and Human Behavior, 647
SDMT. *Véase Symbol Digit Modalities Test*
Seashore Rythm Test (Prueba Rítmica Seashore), 398
Semejanzas (subprueba Wechsler), 220 a 221
Sensibilidad, 422
Sequential Test of Educational Progress (Pruebas secuenciales de progreso educativo), 350, 355
Serie de respuestas de deseabilidad social, 618
Sesgo de prueba, 290 a 302
definición de, 290, 292 a 293
diferencias raciales y, 291
justicia de prueba y, 300 a 302
y validez,
de constructo, 297 a 299
de contenido, 293 a 294
predictiva, 294
Shipley Institute of Living Scale, 249 a 251
Short Portable Mental Status Questionnaire (Cuestionario Portátil Breve del Estado Mental), 417
SIB-R. *Véase Scales of Independent Behavior-Revised*
SII. *Véase Strong Interest Inventory*
SILS. *Véase Shipley Institute of Living Scale*
Símbolos (subprueba Wechsler)
búsqueda de, 225
y dígitos, 224 a 225

Simulación, 453 a 455
Síndrome alcohólico fetal, 307
SIRS. *Véase Structured Interview of Reported Symptoms*
Social Relations Survey (Encuesta de Relaciones Sociales), 550, 551
Social Skills Rating System (Sistema de Calificación de Habilidades Sociales), 444
Speech Sounds Perception Test (Prueba de Percepción de Sonidos del Habla), 398
Spiritual Experience Index (Índice de la experiencia espiritual), 572 a 573
Spiritual Well-Being Scale (Escala de Bienestar Espiritual), 568 a 570
Standards for Educational and Psychological Testing, 52
Stanford Achievement Test (Prueba de Ejecución Stanford), 28, 355
Stanford-Binet, 23
 Cuarta Edición, 198 a 199, 232 a 238
STEP-III. *Véase Sequential Test of Educational Progress*
Strong Interest Inventory (Inventario Strong de Intereses), 529 a 532
Strong Vocational Interest Blank (Formulario Strong de Intereses Vocacionales), 32, 529
Structured Interview of Reported Symptoms (Entrevista Estructurada de Síntomas Reportados), 454 a 455
Subtracting Serial Sevens (Resta de Series de Siete), 397
Sucesión de Letras y Números (subprueba Wechsler), 221
Superyó, 578
SVIB. *Véase Strong Vocational Interest Blank*
Symbol Digit Modalities Test (Prueba de Modalidades de Símbolos y Dígitos), 397
Szondi Test, 31

T

Tabla(s)
 de especificaciones, 153 a 154
 Taylor-Russell, 129 a 131
TAI. *Véase* Test de Apercepción Infantil
Talentos especiales, 445 a 448
TAP. *Véase Tests of Achievement and Proficiency*
Tarea de Frases Incompletas, 602
Tasa base, 129
TAT. *Véase* Test de Apercepción Temática de Thompson, 608
TC. *Véase* Tomografía computarizada
TDAH. *Véase* Trastorno por déficit de atención con hiperactividad
Técnica
 de Completamiento de Actividad, 602
 del dibujo proyectivo de Casa-Árbol-Persona (H-T-P), 31, 611 a 612
 Q, 583 a 584
TEMAS, 609
Teoría(s)
 clásica del error de medición, 94 a 96
 de decisión, 126
 de la capacidad para la generalización, 107 a 109
 de las inteligencias múltiples, 185 a 186
 fenomenológicas, 583 a 585
 triárquica de la inteligencia, 186 a 189
TEP. *Véase* Tomografía por emisión de positrones
Terapia conductual, 647
 cognitiva, 650
Teratógenos, 307
Test(s)
 Breve de Inteligencia de Kaufman, 240 a 241
 Critiques, 48
 de Apercepción,
 Infantil, 607 a 608
 Temática, 31, 604 a 606
 de Aptitudes Diferenciales, 330 a 333
 de Laberintos Porteus, 22, 407
 del Dibujo de la Figura Humana (Goodenough-Harris), 269 a 271
 Guestáltico Visomotor, 404
 No Verbal de la Universidad de Purdue, 410, 498 a 499
 of Achievement and Proficiency (Pruebas Vocacionales y de Aprovechamiento), 354
 of Everyday Attention (Prueba de Atención Cotidiana), 395
 of General Educational Development (Pruebas de Desarrollo Educativo General), 354 a 355
Testigo experto, 452 a 453, 472 a 474
The Future of an Illusion (El porvenir de una ilusión), 564
The Varieties of Religious Experience, 565
Tiempo de reacción, 8 a 9
Time Urgency and Perpetual Activation Scale (Escala de Premura y Actividad Continua), 621
Tinkertoy Test (Prueba Tinkertoy), 407 a 409
T.M.R. School Competency Scales (Escalas de Competencia Escolar T.M.R.), 285
Tomografía
 computarizada, 387
 por emisión de positrones, 387
TONI-3. *Véase* Prueba de Inteligencia No Verbal-3
Trastorno por déficit de atención con hiperactividad, 438

V

VABS. *Véase Vineland Adaptive Behavior Scales*
Validación
 cruzada de pruebas, 162
 ilusoria, 612
Validez, 117 a 140
 ampliación del espectro de la, 137 a 140
 coeficiente de, 125 a 126
 concurrente, 122, 124
 convergente, 135
 de constructo, 131 a 137
 de contenido, 119 a 122
 de discriminación, 135
 de face, 122
 predictiva, 125 a 126
 preocupaciones externas a la, 137 a 140
 reducción de la, 162 a 163
 relacionada con el criterio, 122 a 126
Valor, 555
Valoración
 de custodia de los niños, 464 a 469
 del desempeño, 507 a 516
Valores laborales, 544
Varianza, 74
Ventrículos, 366
Vineland Adaptive Behavior Scales (Escalas Vineland de Conducta Adaptativa), 285

Violencia, predicción de la, 462 a 464
Vocabulario (subprueba Wechsler), 219
Vocational Preference Inventory (Inventario de Preferencia Vocacional), 534 a 537
VPI. *Véase Vocational Preference Inventory*

W

WAIS-III. *Véase* Escala Wechsler de Inteligencia para Adultos-III
Washington University Sentence Completion Test (Prueba de oraciones incompletas de la Universidad de Washington), 602
Watson contra el *Fort Worth Bank and Trust*, 522 a 523
Wechsler Individual Achievement Test (Prueba Wechsler de Rendimiento Individual), 435, 436
Wechsler Memory Scale-III (Escala Wechsler de Memoria-III), 399 a 400
Western Aphasia Battery (Batería Occidental de Afasia), 404
WIAT. *Véase Wechsler Individual Achievement Test*
Wide Range Achievement Test-III (Prueba de Rendimiento de Amplio Rango-III), 435, 436
WISC-III. *Véase* Escala Wechsler de Inteligencia para el Nivel Escolar-III
Wisconsin Card Sorting Test (Prueba Wisconsin de Clasificación de Tarjetas), 407
WJ-R. *Véase* Batería Psicoeducativa Woodcock-Johnson-Revisada
WMS-III. *Véase Wechsler Memory Scale-III*
Work Values Inventory (Inventario de Valores Laborales), 546 a 547
Working, 541
WPPSI-R. *Véase* Escala Wechsler de Inteligencia para los Niveles Preescolar y Primaria-Revisada
WRAT-III. *Véase Wide Range Achievement Test-III*

Esta obra ha sido publicada por
Editorial El Manual Moderno, S.A. de C.V.,
y se han terminado los trabajos de esta
octava reimpresión de la primera edición
el 7 de diciembre 2012,
en los talleres de
Fuentes Impresores, S.A. de C.V.,
Centeno No. 109,
Col. Granjas Esmeralda, 09810.
México, D.F.

8a. reimpresión, 2012